An Interface Design Catalog for Interactive Labeling Systems

Lucas Viana[1][a], Edson Oliveira[2][b] and Tayana Conte[1][c]

[1] *Department of Computing, Universidade Federal do Amazonas (UFAM), Amazonas, Brazil*

[2] *Secretaria de Estado da Fazenda (SEFAZ), Amazonas, Brazil*

Keywords: Interactive Labeling, Interactive Machine Learning, Interface Design, Interface Design for Interactive Machine Learning, Training Data.

Abstract: Machine Learning (ML) systems have been widely used in recent years in different areas of human knowledge. However, to achieve highly accurate ML systems, it is necessary to train the ML model with data carefully labeled by specialists in the problem domain. In the context of ML and Human Computer Interaction (HCI), there are studies that propose interfaces that facilitate interactive labeling by domain specialists, in order to minimize effort and maximize productivity. This paper extends a previous secondary study that discusses some labeling systems. This paper proposes a catalog of design elements for the interface development of this type of system. We built the catalog based on the interface elements found in the studies analyzed in the previous secondary study. With this contribution, we expect to improve the development of better interfaces for interactive labeling systems and, thus, enhance the development of more accurate ML systems.

1 INTRODUCTION

Machine Learning (ML) technology has been successfully applied to different problems from various domains of human knowledge: from art to medicine (Gillies et al., 2016). The advances obtained with the use of this technology, have made considerable impacts on our daily lives (Qiu et al., 2016; Rudin and Wagstaff, 2014). As such, one of the examples is the use of ML for the construction of textual summaries about live events from the status updates of Twitter users (Nichols et al., 2012).

The success of ML projects, however, hides the considerable human work involved in the process of its development (Fiebrink and Gillies, 2018). Developers of ML projects should design a solution considering what is possible to be trained, obtain the necessary data and continuously adjust the learning algorithms according to the errors presented, until an acceptable model for use in practice is achieved. Nevertheless, to obtain this data means to obtain a database properly prepared for the ML development. Thus, it is necessary that this database be representative and accurate for the domain in question, and this is not always easy to obtain (Yimam et al., 2015; Benato et al., 2018).

If it is not possible to obtain a database already prepared for the problem to be solved with ML, the only alternative is to prepare a new database, during the ML development. To prepare this database, it is necessary to involve specialists in the domain in which ML is being applied. These specialists will help by creating and labeling the data, as well as correcting any errors found during this process. The data labeling provides the transfer of knowledge from the domain specialists to the ML model. However, data labeling is a costly, error-prone and labor-intensive process that can frustrate specialists (Nadj et al., 2020). In addition, if the specialist gets frustrated or tired, it can compromise the quality of the data labeling.

Some studies (Thomaz and Breazeal, 2008; Bryan and Mysore, 2013; Self et al., 2016; Nadj et al., 2020) have investigated the use of interactive systems that can enable and facilitate the data labeling, thereby mitigating the frustration of specialists during this process. Nadj et al. (2020) present a Systematic Literature Review (SLR) focusing on interactive labeling systems, in which 44 papers were analyzed. As a result, the authors identified and discussed five design principles for interactive labeling systems. They affirmed that the development and design of these interactive labeling systems require good visual User In-

[a] https://orcid.org/0000-0002-8154-0062
[b] https://orcid.org/0000-0001-9168-4388
[c] https://orcid.org/0000-0001-6436-3773

Viana, L., Oliveira, E. and Conte, T.
An Interface Design Catalog for Interactive Labeling Systems.
DOI: 10.5220/0010458204830494
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 2*, pages 483-494
ISBN: 978-989-758-509-8

terface (UI) planning, since users are not often happy to be treated as mere oracles.

However, Nadj et al. (2020) did not present possible designs of interface elements for interactive labeling systems in their design principles. The choice of interface elements (UI design) is a form of implicit communication between the UI designer and the user (De Souza, 2005). Depending on this choice, the quality of UI interaction can lead to the problems of fatigue and frustration of the users. In this paper, we extend the work of Nadj et al. (2020) to answer the following research question not yet covered in the study "Which interface elements have been used in the interactive labeling systems in the literature?". For this, we grouped the labeling systems found in the studies, extracted UI designs (with their interface elements), creating a catalog to assist developers and designers of new interactive labeling systems.

This paper is organized as follows: in Section 2, the concepts of the theoretical framework are presented; in Section 3, the methodology we followed to perform this analysis in the format of an SLR extension; in Section 4, the results obtained; in Section 5, discussions and, finally, Section 6 presents the conclusions of this paper.

2 THEORETICAL FRAMEWORK

In this section, we introduce the concept of ML, Interactive Labeling, and Interactive Machine Learning (IML). The concept of ML is important to better understand the context in which Interactive Labeling is inserted. The IML concept is important for understanding the analysis of the interfaces performed on the results.

2.1 Machine Learning

ML is a subarea of Artificial Intelligence that focuses on pattern recognition from existing data (Alpaydin, 2020). With pattern recognition of data, ML can make predictions about new data (Mohri et al., 2018). An example of an ML application is the system developed by Nichols et al. (2012). The researchers developed an ML algorithm that builds a journalistic summary of events. The authors Nichols et al. (2012) used Twitter status updates as a source. The objective of the authors is to provide a summary for people, who are not following a live event, to inform themselves quickly about the event. In this case, as new data about a live event appears on Twitter, the ML model predicts a new summary.

ML algorithms can be classified into different learning methods (Zhang, 2020). In the next subsections, we describe the best known learning methods of ML algorithms: reinforcement learning, unsupervised learning, and supervised learning.

2.1.1 Reinforcement Learning

Reinforcement Learning is a learning method where the algorithm transforms feedback on actions taken into learning (Alpaydin, 2020). The purpose of the algorithm is to learn by interacting with the environment. It maximizes its learning as it receives feedback from its actions. These feedbacks can be of the reward type, if the action was correct, or of the punishment type, if the action was not correct. For example, a robot that can receive feedback on its actions and learn from the mistakes and successes (Thomaz and Breazeal, 2008).

2.1.2 Unsupervised Machine Learning

Unsupervised learning method algorithms learn how to categorize unlabeled data into clusters based on data similarity characteristics (Mohri et al., 2018). However, the generated clusters need to be analyzed to see if they make sense within the context of the problem.

Kim et al. (2015) present a study that applied unsupervised ML. The authors presented the benefits generated by using ML to correct activities in a Massive Open Online Course (MOOCs). The ML algorithm created groups of students' activities according to the common characteristics of each activity. By grouping the students' activities with ML, it was possible for teachers to correct once a group of similar activities. This allowed for a reduction in the workload in correcting activities, in addition to allowing the teacher to provide feedback to students more quickly.

2.1.3 Supervised Machine Learning

Supervised Machine Learning (SML) is the learning method in which the ML algorithm learns from a labeled sample database. The labeled sample database is intended to help a learning algorithm to learn and later be able to predict theses labels from new data samples (Kotsiantis et al., 2007).

SML algorithms achieve good performance when the sample database is representative of the problem domain, i.e., the sample database has a good diversity with sufficient quality and quantity (Cui et al., 2014). In other words, SML algorithms are sensitive to the labeled samples to which they are exposed, that is, if the samples are labeled wrong, the algorithms will learn incorrectly, a situation known as “garbage in, garbage out” (Geiger et al., 2020). However, getting a good database of labeled samples for generating an accurate ML model is not an easy task.

2.2 Interactive Labeling

The concept of interactive labeling brings the expert user, i.e., the domain specialists, into the development of the ML model, specially in the SML method, in which some benefits generated are the reduction of time and cost for acquisition of labeled samples and that data samples are labeled under evaluation of the knowledge of the expert user (Cui et al., 2014; Yimam et al., 2015). Ware et al. (2001) observed that, with an easy and intuitive interface, expert users in the field of data knowledge were able to help in the creation of good classifiers (i.e., in the ML model) after a short period of practice with the systems. In addition, when the interactive labeling system is well integrated with other systems, the ML model in question will always be fostered, learning new concepts and even updating existing concepts (Huang et al., 2013).

2.3 Interactive Machine Learning

The ML methods that the ML developer wants to use in her project must be chosen carefully, as this choice affects how the data should be structured and what information that data should contain. In addition, the ML developer must know where to get the data, one of the ways is through specialist in the problem domain.

The interaction between specialists and the developing ML model was traditionally accomplished through the exchange of electronic files, such as spreadsheets or text documents. Ware et al. (2001) and Fails and Olsen Jr (2003) were the pioneers in proposing an interface system with the specialists to facilitate this interaction, calling it Interactive Machine Learning – IML. Their studies presented the benefits caused by the interaction of specialists users of the IML system, in contrast to the traditional practice of interaction through electronic files.

Fails and Olsen Jr (2003) stated that the IML approach breaks with several premises of traditional ML development. According to the authors, with IML it is possible to reduce the chances of the generation of a limited ML knowledge base, in which the generated ML model has to achieve excellent performance when applied to new data never seen in its development. However, according to Dudley and Kristensson (2018), the design of an appropriate interface is fundamental in order to obtain good results with IML systems and this presents a challenge for the design of the interface.

IML systems therefore play an important role in enabling knowledge transfer from specialist users for the development of a successful ML model Behringer et al. (2017). This knowledge transfer occurs through interface elements that must be in the language of these users, who mostly do not have technical knowledge about ML. This enables specialist users tasks related to inspection and correction of ML models to be performed intuitively and effectively (Dudley and Kristensson, 2018).

2.3.1 IML Interface Communication Channels

In the context of IML systems, Dudley and Kristensson (2018) present four interface communication channels, which refer to how the user can interact with these systems. These communication channels define the interaction between the specialist user, the ML model, and the data. These four communication channels idealized by Dudley and Kristensson (2018) are presented below.

Sample Review. This is the way in which the IML system presents the data in the interface to the specialist user in order for her to understand these. In this communication channel, it is important that designers reflect on the following: How should the data be presented to the user in an easy and intuitive way? What can be highlighted from the data in order to facilitate the comprehension and understanding of the specialist user?

Feedback Assignment. This is the way in which the IML system allows the user to perform feedback on the data presented. In this communication channel, designers should reflect on what information from the data and the ML model in development should the IML system offer possible feedback on, and how the interface can be organized and what it should be composed of in order to capture this feedback.

Model Inspection. In this channel, the IML system presents the overall performance of the ML model learned by the machine, and allows the user to check whether the ML model is ready for use, i.e., whether the model is of good quality or if it needs improvements. In this communication channel, it is important that designers reflect on what characteristics and metrics of the ML model already learned should be presented and how these characteristics and metrics should be presented without many ML technical details.

Task Overview. This is the way that the IML system informs the specialist user about the overall progress of the development of the ML model, for example, presenting how much data has already been reviewed and how much data can still be evaluated for improvement of the ML model. In this communication channel, designers should reflect on what information about the overall vision of the task should be presented to the specialist user and how should it be presented?

3 METHODOLOGY

A SLR seeks to identify, select, evaluate, summarize, and interpret phenomena of interest (Kitchenham and Charters, 2007). Also, SLR is documented through a protocol and must allow its reproduction by other researchers (Madeyski and Kitchenham, 2015). Another point is that the SLR can also be extended and updated (Rivero et al., 2013). The extension can be done by asking other research questions that were not asked in the base SLR or an update of the SLR can be carried out, including studies from other years, libraries, updating of new terms in the strings, etc. that were not used in the base SLR.

Based on exploratory searches in the literature, we found the SLR of Nadj et al. (2020), recent and aligned with the context and objective of our study: interactive labeling systems. However, the questions we wish to answer were different. Therefore, we decided to inquire the same set of articles investigated by Nadj et al. (2020) and apply our research questions.

The research question that was addressed in this research is:

- Which interface elements have been used in the interactive labeling systems in the literature?

The following are the steps carried out during the extension of the SLR: selection criteria, selection of studies, data extraction strategy, data extraction, conduction strategy and results.

3.1 Study Selection Strategy

Nadj et al. (2020) found a total of 44 studies that present interactive labeling systems. However, not all these studies describe or present their UI design. As an example, one study was focused on investigating the effects of active human-controlled learning on robots (Cakmak et al., 2010); another study focused on investigating strategies to obtain meaningful samples for labeling (Zhu and Yang, 2019). These studies do not have sufficient details about their interface elements, therefore they do not contribute to answer of our research question. Due to this, we applied a selection criteria to filter whether a study will be included or excluded from our study.

The criteria applied were as follows:

- Inclusion Criteria: the study presents interface elements for interactive labeling systems;
- Exclusion Criteria: the study does not present interface elements for interactive labeling systems.

3.2 Data Extraction Strategy

We conducted the data extraction from the selected studies by analyzing the labeling system interfaces available in the studies. The selected studies were read completely by one of the authors, that also extracted the interface elements. The other authors revised all the extractions results.

In addition to the system interfaces, we also extracted other data:

- Title, authors, year of publication, publication vehicle. This data was collected in order to analyze the timeline of the articles and the most relevant publication vehicles in the area.
- Type of labeling system. This data represents the general objective in which the labeling system is being applied in the analyzed study.
- Problem domain. This data informs what activity the interactive labeling system requests the user during the labeling process.
- Area of application of the labeling system. This data informs which context of the use of the labeling system.
- Interface elements. These data were collected to capture the interface elements of the interactive labeling systems that were not presented visually in the studies, that is, through a graphical interface.

4 RESULTS

This section reports the results of our SLR extension. Subsection 4.1 present the selected studies. Subsection 4.2 presents the results of our research questions.

4.1 Selected Studies

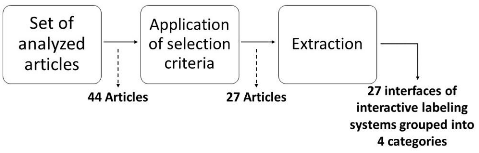

Set of analyzed articles
Application of selection criteria
Extraction
44 Articles
27 Articles
27 interfaces of interactive labeling systems grouped into 4 categories

Figure 1: The methodology used for data extraction and analysis.

Nadj et al. (2020) found a total of 44 studies in their SLR. Out of these 44 studies, we excluded 17 papers because of our selection criteria. Therefore, we selected 27 papers that presented interface elements, as shown in Figure 1. Of these 27 studies, 9 were published in journals, 13 in conferences, 4 in workshops and 1 paper was related to a technical report.

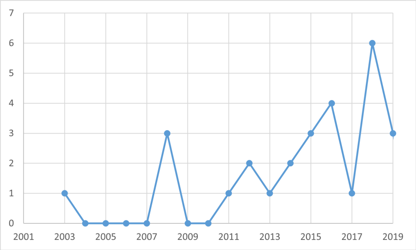


Figure 2: Period of publication of the articles.

As shown in Figure 2, the selected 27 studies were published from 2003 to 2019. The year 2018 had more publications in comparison with the other years, with a total of 6 publications. Also, it is possible to identify a growing trend in interest in interactive labeling systems. We identified 4 types of data processed in the interactive labeling systems. As shown in Figure 3, we classified the data type into sound, image, textual, and generic.

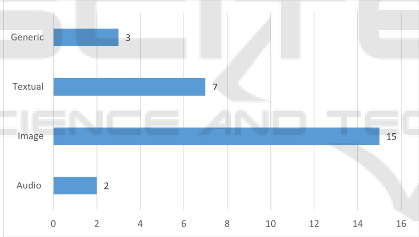


Figure 3: Data Type of Labeling Systems.

4.2 RQ: Which Interface Elements Have Been used in the Interactive Labeling Systems in the Literature?

To answer our research question, we divided the analysis into two parts. The first part is the analysis of clusters of the analyzed interactive labeling systems. The second part refers to the composition of the interfaces.

4.2.1 Groups of Interactive Labeling Systems

For this analysis, we consider the different approaches applied to soliciting user labeling. We were able to find 4 groups shown in Table 1.

The first group is composed by the systems that expect direct labeling on a piece of data or a set of data from the user, which we named “direct labeling approach”. In this grouping, the systems request direct labeling from the user regarding the category of the data, as in the study of Fogarty et al. (2008). In this first group, the systems usually ask the user to choose between “correct” or “wrong” regarding an object label, to evaluate whether it is in its correct category or not. However, some of these systems do not request user labeling in only two categories. For example, the abstrackr system from Wallace et al. (2012) goes beyond providing 3 alternatives for the user: “correct” or “wrong”, or “maybe”.

Some systems differ in the labeling information that is requested from the user. The labeling systems presented in Guo et al. (2018); Xu et al. (2017), ask the user for feedback through comparison with other data. We call this group the “indirect labeling approach”.

There are also interactive labeling systems that solicit user labeling in an exploratory way, which we call the “exploratory labeling approach”. This type of strategy is applied in tasks where the user must identify the objects and their limits in an image. In this approach, all the interactive labeling systems analyzed belong to the supervised ML method.

Finally, the UI’s classified in the “N/A ” group were those that the system did not ask for the data class. In this group are systems such as the Self et al. (2016) that requests feedback regarding the relevance of the data attributes.

4.2.2 Analysis of the Composition of the Interfaces

The results in this analysis consider the interface screens that were presented in the studies. To facilitate the organization of UIs and their elements, we have created a catalog for this. The catalog was built from the perspective of the communication channels identified by Dudley and Kristensson (2018).

Figure 4 presents the relationship between the composition of interactive labeling system interfaces and communication channels by Dudley and Kristensson (2018). All the 27 analyzed system interfaces present elements that were classified as the sample review communication channel. In 23 of the interfaces, there are interface elements that configure the feedback assignment communication channel. Only one interface has elements that configure the model inspection communication channel. Finally, 2 interfaces have elements that configure the task overview communication channel.

Of the 27 systems analyzed, only 2 of these are focused on assigning a graphical interface for the construction of an ML model. This graphical interface

Table 1: Groups of interactive labeling systems.

Groups	Articles
Direct labeling approach	Weigl et al. (2016); Zhang et al. (2008); Burkovski et al. (2011); Fogarty et al. (2008); MacGlashan et al. (2017); Cheng et al. (2016); Kim et al. (2015); Wallace et al. (2012); Amershi et al. (2014); Yimam et al. (2016).
Indirect labeling approach	Guo et al. (2018); Xu et al. (2017); Plummer et al. (2019); Amershi et al. (2014).
Exploratory labeling approach	Bryan and Mysore (2013); Jain et al. (2019); Fails and Olsen Jr (2003); Nalisnik et al. (2015); Kim and Pardo (2018); Boyko and Funkhouser (2014); Acuna et al. (2018); Harvey and Porter (2016).
N/A	Dasgupta et al. (2019); Self et al. (2016); Bernard et al. (2017); Amershi et al. (2012); Thomaz and Breazeal (2008); Amershi et al. (2014); Datta and Adar (2018).

makes it easy for ML professionals to use it as well as for users interested in using ML without technical knowledge for data manipulation. The other 25 systems are focused on creating a system for the user to perform data labeling.

The systems that are for use in the audio area are intended for the labeling of sounds in sound recordings through spectrogram. One example is the system presented by Kim and Pardo (2018), shown in Figure 5-B. Image systems are those that treat the identification of objects in an image through user labeling, such as the study of Jain et al. (2019). For the systems that act in the textual area, one example is the study of Yimam et al. (2015), it aims to identify knowledge of the medical area through reports of medical diagnoses, shown in Figure 6-D. Finally, the generic systems were those that did not fit into the previous areas. Figure 3 shows the number of systems analyzed by area of activity.

The 27 system interfaces analyzed present many ideas for implementing interface elements. However, due to the number of design elements found, only a few of these will be shown below in Figures 5 and 6. The complete list of interfaces, their elements and the classification in the communication channels of Dudley and Kristensson is available at: https://figshare.com/articles/figure/_/13785727.

The interface elements that characterize the communication channels of Dudley and Kristensson (2018) are highlighted in a red box with a number by the side (see Figures 5 and 6). The number represents the identifier of the communication channels. The identifiers are 1 - Sample Review, 2 - Feedback Assigment, 3 - Model Inspection and 4 - Task Overview.

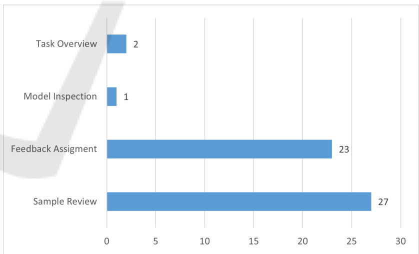
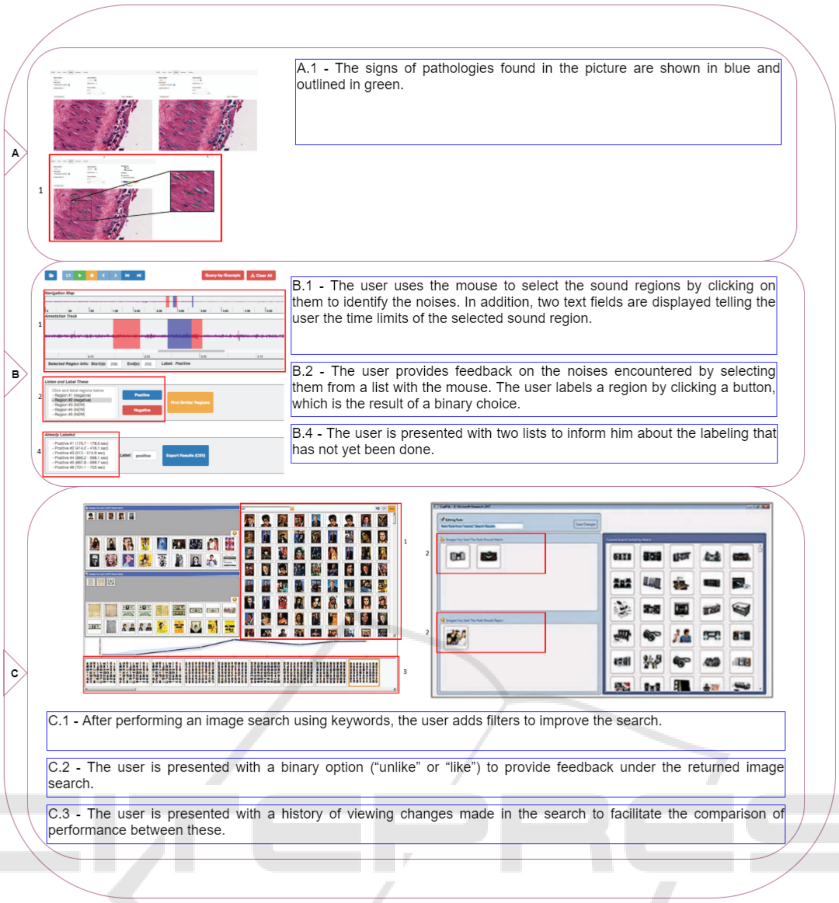


Figure 4: Relationship between the interface elements of the analyzed systems and the communication channels of Dudley and Kristensson (2018).

Most interfaces do not have all communication channels. Therefore, not all interfaces contain the four communication channels identified in Figures 5 and 6. Each letter shown in Figures 5 and 6 (A, B, C, D, and E) identifies a different labeling system. The authors of these studies are referenced in the caption of Figure 5 and 6 according to their identifying letter.

The interface presented with the identifying letter “A” in Figure 5 is a system that aims to identify pathologies in medical imaging examinations (Nalisnik et al., 2015). For this purpose, the users of the



A.1 - The signs of pathologies found in the picture are shown in blue and outlined in green.

B.1 - The user uses the mouse to select the sound regions by clicking on them to identify the noises. In addition, two text fields are displayed telling the user the time limits of the selected sound region.

B.2 - The user provides feedback on the noises encountered by selecting them from a list with the mouse. The user labels a region by clicking a button, which is the result of a binary choice.

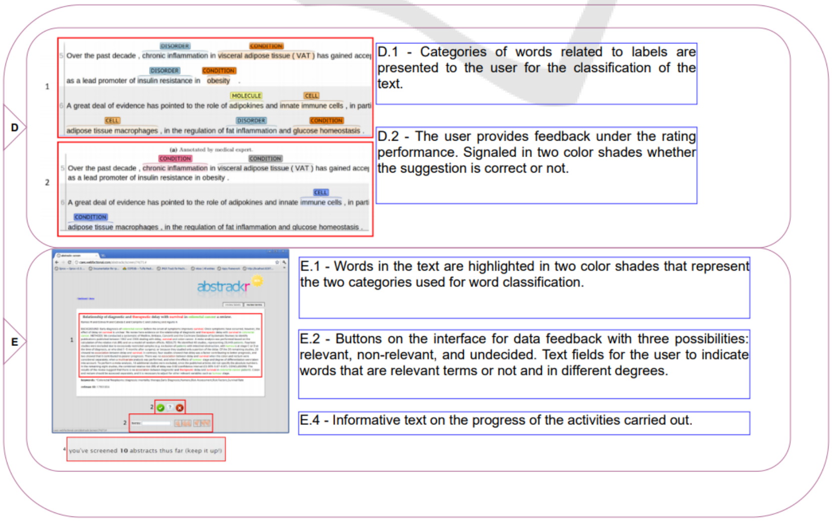
B.4 - The user is presented with two lists to inform him about the labeling that has not yet been done.

C.1 - After performing an image search using keywords, the user adds filters to improve the search.

C.2 - The user is presented with a binary option (“unlike” or “like”) to provide feedback under the returned image search.

C.3 - The user is presented with a history of viewing changes made in the search to facilitate the comparison of performance between these.

Figure 5: Reduced catalog - Part I: consisting of 3 labeling systems presenting interface elements characterizing the communication channels. The interfaces used in this catalog were captured from studies by the authors Nalisnik et al. (2015) in (A), Kim and Pardo (2018) in (B), Amershi et al. (2014) in (C).



D.1 - Categories of words related to labels are presented to the user for the classification of the text.

D.2 - The user provides feedback under the rating performance. Signaled in two color shades whether the suggestion is correct or not.

E.1 - Words in the text are highlighted in two color shades that represent the two categories used for word classification.

E.2 - Buttons on the interface for data feedback with three possibilities: relevant, non-relevant, and undecided. Text fields for the user to indicate words that are relevant terms or not and in different degrees.

E.4 - Informative text on the progress of the activities carried out.

Figure 6: Reduced catalog - Part II: consisting of 2 labeling systems presenting interface elements characterizing the communication channels. The interfaces used in this catalog were captured from studies by the authors Yimam et al. (2015) in (D) and, Wallace et al. (2012) in (E).

system are those with experience in image examination reading. They should analyze the images and identify the signs of pathologies by means of a virtual brush that users employ to highlight the signs of pathologies. Through this interaction between the labeling system and the user, a database labeled by the users is built. After this, a training of the algorithm for automatic identification in new images of exams was carried out.

The system interface presented with the identifier letter "B" in Figure 5 focuses on the recognition of errors in sound recordings (Kim and Pardo, 2018). Users of this system are specialists in visual and sound reading of digital sounds. These users must identify and recognize the areas of the sound recording that are the bad ones. The interface elements involved in this system are the program for analyzing the sound recording, the list of regions of the spectrogram already labeled by the user, and the list of regions to be labeled.

The interface of the system presented with the identifier letter "C" in Figure 5 shows a system that aims to perform searches of images with keywords and rules defined by the user (Amershi et al., 2014). For this system, expert users do not have a specific profile. These users generate the search rules by interacting with the first search they perform by selecting the images returned erroneously, for example.

The interface of the system presented with the identifier letter "D" in Figure 6 presents a system related to identification of medical knowledge by means of annotations. Yimam et al. (2015) demonstrated the impact of the use of an IML system on the development of a database for the recognition of biomedical citations in medical annotations. During the annotation performed by a specialist, an ML model is constructed using as data these annotations in order to propose labels for subsequent annotations. Although the study is still in an exploratory stage at the time of publication, the experimental results indicated qualitatively and quantitatively the feasibility of the method for a more personal and responsive information extraction technology.

Finally, the system interface presented with the identifier letter "E" in Figure 6 presents a labeling system integrated into a common system in the life of users. In this system expert users are most often end users of the system. Wallace et al. (2012) presented a labeling system called "abstrackr". Abstrackr aims to mitigate the workload of medical researchers in the search for evidence (a systematic review of literature) that is pertinent to a specific clinical question. According to the authors, although many ML methods have been proposed and used in the past, they are actually rarely (or never) made available to professionals in practice. Abstrackr has already been used in more than 50 systematic reviews, in which the authors claim to have reduced the workload of medical researchers by 40% without erroneously excluding any relevant reviews.

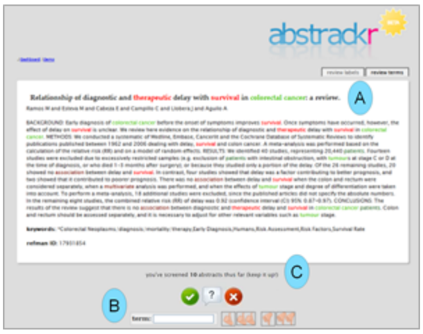

abstrackr
review labels
review terms
A
Relationship of diagnostic and therapeutic delay with survival in colorectal cancer: a review.
you've screened 10 abstracts thus far (keep it up!)
C
B
?
term:

Figure 7: Abstrackr's interactive labeling system - adapted from Wallace et al. (2012).

In order to facilitate the understanding of these interfaces of the labeling systems presented in Figure 5, we will explain more about one of these systems, namely, abstrackr. In Figure 7, we will use as a reference a screen of the abstractkr system, as proposed by Wallace et al. (2012), to further explain the details of this type of interface. The explanation will be conducted by segmenting parts of the interface into the perspectives of communication channels.

4.2.3 Detailed Analysis

In the communication channel "Sample Review", the labeling system presents the data samples to the user, to subsequently obtain from them feedback on the labeling of these. The Abstrackr interface design presents the title and summary of a scientific paper from the medical area (Figure 7-A). Abstractkr's interface elements involve highlighting text words by color, according to relevance or not learned by the machine.

Regarding the communication channel "Feedback Attribution", the labeling system expects interactions from the expert user, given the data sample presented by the "Sample Review", in order to use this feedback in the evolution of the ML model. In the abstractkr tool, the possible interactions by expert users are: text field to provide new terms and classify them as relevant or not, buttons next to the field (Figure 7-B); buttons to label the abstract of the presented paper as relevant (Tick) or not ("X"), or as undecided (white button with a ?).

In the communication channel "Model Inspection", the labeling system presents information about the quality of the ML model already learned by the machine, such as, for example, its current classification accuracy. In accordance with Wallace et al. (2012), the abstrackr tool does not present, at least in what was made available by the paper, in its interface design, any information about the ML model under development. However, this system could have used existing interface elements to represent the communication channel for the user. The user could be informed about the relevance of other abstracts and their assertiveness, or be presented a bar graph with the amount of texts considered relevant or not in the current state of the ML model. This is an example in which the catalog of design elements can assist in the development of labeling systems that better support the expert user in understanding the ML model under construction.

In the communication channel "Task Overview", the labeling system presents the user with information on the progress of the support task in the construction of the ML model. The abstractkr system provides the progress of readings performed by the expert user in the systematic review, as can be seen in Figure 7-C. In addition to the abstractkr system, only one system displays elements in the interface that characterizes the communication channel "Task Overview" and is found in the study of Kim and Pardo (2018). In this case, with the catalog, it was possible to identify the need for further studies focusing on the presentation of this channel in interfaces of labeling systems. With new studies, more paths emerged to provide other possibilities beyond those used in the studies of Wallace et al. (2012) and Kim and Pardo (2018) for the developers and designers of these systems.

5 DISCUSSION

In this study, we identified design ideas for interfaces of interactive labeling systems, also, we found 4 groupings of types of these systems. We classify systems as feedback approach, indirect feedback approach, exploratory feedback approach, and N/A. Finally, we were able to answer our research question through the catalog and the groupings of the systems.

According to Katan et al. (2015), there are few studies that explore interface design for labeling systems. Furthermore, according to the authors, to gain a better understanding of this type of design, including related challenges, it is necessary to explore existing systems and research other alternatives to interface design. In this sense, this study extracted a catalog with twenty-seven user interface design ideas of labeling systems developed in selected studies from literature.

Analyzing our results, one can notice that the choice of interface elements varies according to the strategy for capturing the feedback regarding the labeling. Also, another factor that influences the choice of interface elements is the data type. In interactive labeling systems in which the data type is an image, there are several interface elements that we have identified in the catalog that can be used for new labeling system projects. However, when the data type is textual, the choices of interface elements are limited.

In our view, the interface design of labeling systems for some types of data is more intuitive than other types. For example, in the classification of knowledge data of media, such as photos or videos, the very presentation of the example without any modification facilitates the design ideas to be worked on in the labeling system of Fails and Olsen Jr (2003), such as the use of markers in the images to highlight important parts. Its own image already visually collaborates intuitively for the elaboration of interface elements.

This is also the case in sound labeling systems. In the study of Bryan and Mysore (2013), a labeling system was developed for the identification of noises in sound recordings, however, the sound is in another type of media (the spectrogram). The original midia is represented in another metaphor that helps the user of the labeling system to interact with the system.

In the labeling of textual data, it is important that the interface elements support expert users in a more efficient reading, avoiding textual data that is unnecessary and/or irrelevant, confusing, long, among other problems. These situations related to textual data can cause fatigue and stress in the expert user and, especially, a lack of understanding during their labeling feedback during the use of the system. This is the type of interface design that affects the quality of review of the examples by expert users, and impacts the success or failure of the ML model under development.

In addition to the twenty-seven design ideas extracted and cataloged, other ideas of interaction without a visual design were observed. One was to carry out a double review of the data, making the assessment more assertive when agreed by the two specialists (Wallace et al., 2012). Also, other studies proposed an investigation strategy to the user the degree of intimacy and trust under the data labeling, in addition to internally assessing the impact of the generated labeling (Kim et al., 2015).

The representation of the data that is required to be labeled in the labeling systems often needs to go through the process of metaphor for a type of data different from its original. We observed the proposal for the labeling system presented in Kim and Pardo (2018), the process of transforming one type of data into another brought more possibilities for the use of interface elements and consequently a better interaction with the user.

We believe that the process of transforming one type of data to another can improve the interaction between the user and the interactive labeling systems, providing the user with an improvement in the fluidity of tasks within these systems, preventing the user from feeling unmotivated, compromising quality data labeling.

This research, in addition to providing a catalog, can serve as a guide for exploring aspects of existing ideas for the interface of interactive labeling systems.

6 CONCLUSIONS

Interactive labeling systems are important in various contexts, such as support for health professionals and business professionals. However, there are few studies in the literature that explore the analysis of these systems due to their complexity (Kim et al., 2015). This paper extends the SLR of Nadj et al. (2020) focusing on the analysis of interface elements used to design interactive labeling systems. The outcome is a catalog of interface elements classified from the perspective of communication channels established by Dudley and Kristensson (2018).

The catalog aims at being a helpful tool for the development of labeling systems interfaces, presenting solutions that have proven effective in the scientific literature. Bad UI design choices usually lead to problems of users' fatigue and frustration. Furthermore, the catalog is useful for directing what the interfaces should show the user, in addition to demonstrating practical applications from other systems that serve the guidelines for developers.

One of the limitations of this study is how the analysis of interfaces were conducted. The analysis were carried out with the interface screens available in the analyzed studies. Most of the analyzed systems do not have access for the general public and, therefore, the analysis was limited only to the interface screen and the interface elements described textually in the studies.

With this catalog, we expect to contribute to the effective interaction of specialist users with interactive labeling systems during the ML model construction process. Thus, ML models will be developed with higher quality, and promote a better experience for the end users of these systems.

In future work, we intend to investigate which methods are proposed in the literature to assess the quality of the interactive labeling systems' interface. Exploring these methods can also improve the interface design ideas that exist in the catalog. Besides, we highlight the importance of more research investigating the interaction between these systems and the users using the HCI lens.

ACKNOWLEDGMENTS

We would like to thank the Foundation for Research Support of the State of Amazonas (FAPEAM) for the financial support through POSGRAD and the processes numbers 062.00638/2019 and 062.00150/2020, CNPq through the process number 311494/2017-0, and the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001. Furthermore, this research, according to Article 48 of Decree nº 6.008/2006, was partially funded by Samsung Electronics of Amazonia Ltda, under the terms of Federal Law nº 8.387/1991, through agreement nº 003/2019, signed with ICOMP/UFAM.

REFERENCES

Acuna, D., Ling, H., Kar, A., and Fidler, S. (2018). Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 859–868.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Amershi, S., Fogarty, J., and Weld, D. (2012). Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30.

Behringer, M., Hirmer, P., and Mitschang, B. (2017). Towards interactive data processing and analytics-putting the human in the center of the loop. In *International Conference on Enterprise Information Systems*, volume 2, pages 87–96. SCITEPRESS.

Benato, B. C., Telea, A. C., and Falcão, A. X. (2018). Semi-supervised learning with interactive label propagation guided by feature space projections. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 392–399. IEEE.

Bernard, J., Hutter, M., Zeppelzauer, M., Fellner, D., and Sedlmair, M. (2017). Comparing visual-interactive labeling with active learning: An experimental study.

IEEE transactions on visualization and computer graphics, 24(1):298–308.

Boyko, A. and Funkhouser, T. (2014). Cheaper by the dozen: Group annotation of 3d data. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 33–42.

Bryan, N. and Mysore, G. (2013). An efficient posterior regularized latent variable model for interactive sound source separation. In *International Conference on Machine Learning*, pages 208–216.

Burkovski, A., Kessler, W., Heidemann, G., Kobdani, H., and Schütze, H. (2011). Self organizing maps in nlp: Exploration of coreference feature space. In *International Workshop on Self-Organizing Maps*, pages 228–237. Springer.

Cakmak, M., Chao, C., and Thomaz, A. L. (2010). Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118.

Cheng, T.-Y., Lin, G., Gong, X., Liu, K.-J., and Wu, S.-H. (2016). Learning user perceived clusters with feature-level supervision. In *NIPS*.

Cui, S., Dumitru, C. O., and Datcu, M. (2014). Semantic annotation in earth observation based on active learning. *International Journal of Image and Data Fusion*, 5(2):152–174.

Dasgupta, S., Poulis, S., and Tosh, C. (2019). Interactive topic modeling with anchor words. *Workshop on Human in the Loop Learning (HILL 2019)*.

Datta, S. and Adar, E. (2018). Communitydiff: visualizing community clustering algorithms. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(1):1–34.

De Souza, C. S. (2005). *The semiotic engineering of human-computer interaction*. MIT press.

Dudley, J. J. and Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37.

Fails, J. A. and Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45.

Fiebrink, R. and Gillies, M. (2018). Introduction to the special issue on human-centered machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):7.

Fogarty, J., Tan, D., Kapoor, A., and Winder, S. (2008). Cueflik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 29–38.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. (2020). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Heloir, A., Nunnari, F., Mackay, W., Amershi, S., Lee, B., et al. (2016). Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3558–3565.

Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., and Feris, R. (2018). Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31:678–688.

Harvey, N. and Porter, R. (2016). User-driven sampling strategies in image exploitation. *Information Visualization*, 15(1):64–74.

Huang, S.-W., Tu, P.-F., Fu, W.-T., and Amanzadeh, M. (2013). Leveraging the crowd to improve feature-sentiment analysis of user reviews. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 3–14.

Jain, S., Munukutla, S., and Held, D. (2019). Few-shot point cloud region annotation with human in the loop. *arXiv preprint arXiv:1906.04409*.

Katan, S., Grierson, M., and Fiebrink, R. (2015). Using interactive machine learning to support interface development through workshops with disabled people. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 251–254.

Kim, B., Glassman, E., Johnson, B., and Shah, J. (2015). ibcm: Interactive bayesian case model empowering humans via intuitive interaction.

Kim, B. and Pardo, B. (2018). A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–23.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Roberts, D., Taylor, M. E., and Littman, M. L. (2017). Interactive learning from policy-dependent human feedback. *34th International Conference on Machine Learning (ICML 2017)*.

Madeyski, L. and Kitchenham, B. (2015). Reproducible research–what, why and how. *Wroclaw University of Technology, PRE W*, 8.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Nadj, M., Knaeble, M., Li, M. X., and Maedche, A. (2020). Power to the oracle? design principles for interactive labeling systems in machine learning. *KI-Künstliche Intelligenz*, pages 1–12.

Nalisnik, M., Gutman, D. A., Kong, J., and Cooper, L. A. (2015). An interactive learning framework for scalable classification of pathology images. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 928–935. IEEE.

Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198.

Plummer, B., Kiapour, H., Zheng, S., and Piramuthu, R. (2019). Give me a hint! navigating image databases using human-in-the-loop feedback. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2048–2057. IEEE.

Qiu, J., Wu, Q., Ding, G., Xu, Y., and Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):67.

Rivero, L., Barreto, R., and Conte, T. (2013). Characterizing usability inspection methods through the analysis of a systematic mapping study extension. *CLEI Electronic Journal*, 16(1):12–12.

Rudin, C. and Wagstaff, K. L. (2014). Machine learning for science and society.

Self, J. Z., Vinayagam, R. K., Fry, J., and North, C. (2016). Bridging the gap between user intention and model parameters for human-in-the-loop data analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–6.

Thomaz, A. L. and Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 819–824.

Ware, M., Frank, E., Holmes, G., Hall, M., and Witten, I. H. (2001). Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292.

Weigl, E., Walch, A., Neissl, U., Meyer-Heye, P., Radauer, T., Lughofer, E., Heidl, W., and Eitzinger, C. (2016). Mapview: Graphical data representation for active learning. In *CEUR Workshop Proceedings (2016) 1707 3-8*, pages 3–8.

Xu, Y., Zhang, H., Miller, K., Singh, A., and Dubrawski, A. (2017). Noise-tolerant interactive learning using pairwise comparisons. In *Advances in neural information processing systems*, pages 2431–2440.

Yimam, S. M., Biemann, C., Majnaric, L., Šabanović, Š., and Holzinger, A. (2015). Interactive and iterative annotation for biomedical entity recognition. In *International Conference on Brain Informatics and Health*, pages 347–357. Springer.

Yimam, S. M., Biemann, C., Majnaric, L., Šabanović, Š., and Holzinger, A. (2016). An adaptive annotation approach for biomedical entity and relation recognition. *Brain Informatics*, 3(3):157–168.

Zhang, L., Tong, Y., and Ji, Q. (2008). Active image labeling and its application to facial action labeling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2008) 5303 LNCS(PART 2) 706-719*, pages 706–719. Springer.

Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence*. Springer.

Zhu, Y. and Yang, K. (2019). Tripartite active learning for interactive anomaly discovery. *IEEE Access*, 7:63195–63203.