

# CS 565 - Final Report - CHARLIE

Béatrice Moissinac, Chi Wen, Yi-Jung Chiang, I-Shen Liao

June 8, 2018

# Chapter 1

## Usability Principles

### 1.1 Introduction to the Software

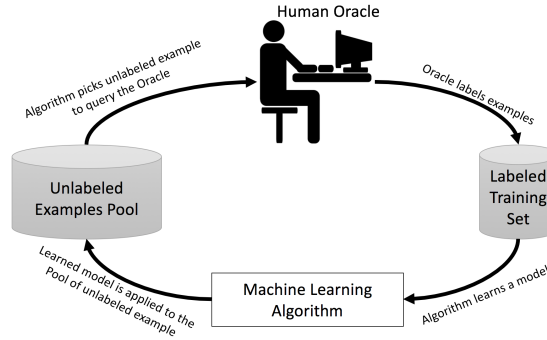


Figure 1.1: Active Learning with Human-in-the-loop

The Software presented in this work is called Curriculum Heuristic for Active Learning with Intent Explanation (CHARLIE). CHARLIE is an Active Learning with human-in-the-loop system which uses different heuristics and explanation to assist the human Oracle in its task. Figure 1.1 describes the overall mechanism. CHARLIE alternates two UI for each of the two tasks for the Oracle: (i) labeling a sequence of 10 instances one at-a-time (Figure 8.1), and (ii) labeling a feature (Figure 8.2). During the instance labeling task (Figure 8.1), the user/Oracle may pick one out of four labels which best describe the sentence (or instance) display at the top of the screen (Figure 8.1b). The user/Oracle also has the option of skipping this example (see Figure 8.1c) in which case a sub-menu appears and presents a way to get feedback on the reason for skipping. In the feature labeling task (Figure 8.2), the user/Oracle may promote feature for each class, based on their own judgment of what feature is relevant for each class. For instance, the feature "beer" may be very relevant to identify an instance of the class "alcohol". The user/Oracle may promote as many feature per round, the promotion persists throughout the session as tasks alternate. A promotion can be removed at any point during the session. The number of feature displayed on the screen starts at 5 words per class and increase by 5 words at each subsequent iteration through this task.

### 1.2 Our Goals

Ultimately, we want to improve the performance of the classifier while minimizing the number of tasks the user/slave/Oracle has to do in order to achieve the desired performance. We posit that this can be done by improving the usability of the current UI. Improving the usability of the UI may improve the consistency of the user/Oracle, as well as engagement, motivation, and quality of the labeling. For instance, our user study has shown that user with more expertise in ML tend to provide better labeling during the 2nd task, because they know of the concept of "split words" and "over-fitting" in classification tasks. We know that they know of these concepts because they mentioned them while thinking at loud during the task.

Below we address our concerns relative to specific design principles.

### 1.2.1 Feedback

Feedback is the main concern for this UI. The UI does not give feedback to the user about the impact of the user's action on the learning agent's performance (BUG001). For instance, how did promoting a feature affect the classification? (BUG002) The learning agent's "performance" is actually a very complex measure. Showing the accuracy rate might not be very helpful to the user to understand what is going on with the classifier. Many Explainable AI (XAI) approaches use complex Machine Learning heuristic that may not be readily interpretable by a non-ML expert. Given an explanation of the classifier (e.g., feature importance as shown in Figure 8.2), what feedback would be the most beneficial to the user's mental model of the learning agent, to help the user help the learning agent? This is a crucial usability question related to the usability of XAI solutions.

### 1.2.2 Constraint

Constraint is another major concern in our UI: the UI is too constrained and affords very little freedom to the user, especially in the instance labeling task. Mainly, the user cannot go back to view (BUG003)/change (BUG004) labels on previously labeled instances, even if it is within the same batch. This is a major flow in the system which render the UI very unsafe.

### 1.2.3 Visibility

There are some issues with visibility in both interfaces of the system. In the first interface (Figure 8.1), it was noted during the user study that participants would rather guess a label rather than skip (BUG007). It is likely that the 'Skip' button is simply outside the vision field of the participant. In the second interface, as the list of features grows, it is not obvious to the user that they need to scroll down. Also, as the list of promoted words stacks to the top, these words are now of little use to select the next "useful" feature for the class (BUG008). As seen in Figure 8.2a, it tends to clutter the page after a while.

### 1.2.4 Usability

Lastly, in addition of the missing feature we mentioned above, we think that the system is missing features allowing the user/Oracle to elicit their mental model of the explanation. A greater part of CHARLIE, beyond the usability question, it to create a tutor agent in between the user and the learning agent, to insure that the user actually understands the explanation, and understands the impact of their action onto the learning agent. In term of usability, we want to investigate how to have the user/Oracle elicit their mental model while maintaining engagement and motivation. This is a hard problem in HCI.

## 1.3 Scenarios

1. Abby wants to check and/or modify a label she has entered during a previous instance labeling iteration (Figure 8.1b).
2. Abby wants to know if her previous feedback to the classifier has had any impact on its performance.
3. Abby wants to save her work and stop working on this.

## 1.4 Persona

Note that we are likely to use a new dataset for this project. The dataset may be movie reviews and the classification would be on the sentiment expressed by the reviews<sup>1</sup>.

---

<sup>1</sup>For instance: <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data>

# Abby Jones<sup>1</sup>



## You can edit anything in blue print

- 24 years old
- She is a student
- Lives in Corvallis, Oregon, USA

Abby has always liked movies. When she has free time, she watches movies on the Internet. She likes a wide spans of genres. But she doesn't like to go to the cinema, she prefers watching from home. She prefers scanning all her emails first to get an overall picture before answering any of them. (This extra pass takes time but seems worth it.) She likes to read movie blogs and read the comments about a movie before watching it.

## Background and skills

Abby is an MBA student. She is comfortable with the technologies she uses regularly to do her homework, but she started an internship somewhere where their software systems are new to her.

Abby says she's a "numbers person" , but she has never taken a Machine Learning or Computational Statistics class. She likes Math and knows how to think with numbers She writes and edits spreadsheet formulas in her work.

In her free time, she also enjoys working with numbers and logic. She likes strategy games, especially online games.

## Motivations and Attitudes

- **Motivations:** Abby uses technologies to accomplish her tasks. She learns new technologies if and when she needs to, but prefers to use methods she is already familiar and comfortable with, to keep her focus on the tasks she cares about.
- **Computer Self-Efficacy:** Abby has low confidence about doing unfamiliar computing tasks. If problems arise with her technology, she often blames herself for these problems. This affects whether and how she will persevere with a task if technology problems have arisen.
- **Attitude toward Risk:** Abby's life is a little complicated and she rarely has spare time. So she is risk averse about using unfamiliar technologies that might need her to spend extra time on them, even if the new features might be relevant. She instead performs tasks using familiar features, because they're more predictable about what she will get from them and how much time they will take.

## How Abby Works with Information and Learns:

- **Information Processing Style:** Abby tends towards a *comprehensive information processing style* when she needs to more information. So, instead of acting upon the first option that seems promising, she gathers information comprehensively to try to form a complete understanding of the problem before trying to solve it. Thus, her style is "burst-y"; first she reads a lot, then she acts on it in a batch of activity.
- **Learning: by Process vs. by Tinkering:** When learning new technology, Abby leans toward process-oriented learning, e.g., tutorials, step-by-step processes, wizards, online how-to videos, etc. She doesn't particularly like learning by tinkering with software (i.e., just trying out new features or commands to see what they do), but when she does tinker, it has positive effects on her understanding of the software.

<sup>1</sup>Abby represents users with motivations/attitudes and information/learning styles similar to hers. For data on females and males similar to and different from Abby, see <http://eusesconsortium.org/gender/gender.php>

## 1.5 Where is Abby?

Abby, or someone like her, will be found among our acquaintance within the OSU and Corvallis community (e.g., friends, colleges in other departments etc.)

## 1.6 Pre-dispositions

EDIT: We assumed it was about Abby.

### 1.6.1 What do we know?

- Abby has no machine learning knowledge.
- Abby can read and understand English
- Abby can use a standard computer unassisted for some specific tasks.
- Abby likes Math and is a "number person".
- Abby likes movies, and likes to read reviews about movies.
- Abby will be biased by her own emotions and interpretation of language during the labeling.
- Abby will gather as much context and information as she can before taking a labeling action.

### 1.6.2 What do we not know?

- Does Abby have any interest or motivation to label movie reviews or helping improve such a classifier?
- Does Abby have a preference toward the type of reviews she read (Does she prefer to read only positive reviews, or only negative reviews?)
- What is the average length of a review Abby usually like to read?
- What is likely to bias her during the process? (Not knowing a word? Length of reviews? Level of language used?)
- Will she be reasonable consistent between sessions if there are several sessions of labeling? Will she remember previous label decisions
- How long would Abby be willing to do this task before losing motivation and engagement?

## Chapter 2

# Heuristic Evaluation

In this chapter, we chose Scenario 2 : “Abby wants to know if her previous feedback to the classifier has had any impact on its performance. For instance, she wants to know if the classifier is now "good enough" for her to stop. ”

### 2.1 Workflow

The researcher starts the experiment by positioning the user in front of the instance labeling screen. Then, the user goes as follow:

1. Read the instance, and the user decides to select a label or to skip.
2. User selects 'Next' or 'Skip' depending
3. User repeats this action 10 times, to label a batch of 10 instances
4. User reads list of words. A chore ;)
5. User selects words to promote
6. User selects 'Continue'
7. User repeat step 1 to 6 many times, until User wants to know how the classifier is doing.

Scenario begins here:

1. User visually scans the “instance labeling” screen, searching for some feedback from the algorithm.
2. User does not find such feedback and has to move on to the next screen, hoping to find more info on the next screen.
3. Repeat 10 times (for the ten examples)
4. User sees the promoting page, scans the page searching for some feedback from the algorithm.
5. User does not find such feedback and has to move on to the next screen, hoping the end is near.
6. Repeat ad nauseam

### 2.2 Workflow Evaluation using the GenderMag Heuristics

#### 2.2.1 Motivations

Abby and Pat will follow the workflow step by step to label examples without problems, but they want to make quick work of these tasks since they don't have a lot of time, so they will quickly want to know how they are doing. When they get to the point when they can't get feedback (BUG001, BUG002), they will quickly give up. Abby will be the first to give up, Pat will persevere a little bit

longer in attempting to understand the system. Tim wants to know how he is doing right away, to adjust his tinkering. He will carry on a long time, even though he realizes that the feedback feature most likely does not exist, then he will quit (BUG001, BUG002).

### **2.2.2 Information-Processing Style**

Abby and Pat will read everything presented to them on the screen. It makes the 2nd task, feature labeling, very tedious, because of the long list of words. Moreover, they both do not know how promoting features affect the performance of the classifier exactly (BUG006). The only information they know is from the short oral tutoring before they started, which stated : " This page provides you a list of words for each label. The classifier thinks that these words are useful to identify a particular label. Your task is to promote words, if you think they are useful to recognize a label. Once you are done, you can select "Continue"." Furthermore, they will be frustrated from not being able to review previously labeled examples (BUG004). Tim made a lot of different labeling decision to tinker with the classifier, until he would see a difference with the data being presented. He is also more likely to miss the 'skip' button because he goes pretty fast through the task. He is worried when he realizes he cannot undo his tinkering (BUG011).

### **2.2.3 Computer Self-Efficacy**

Abby will not persevere with the system for very long. She is not a machine learning and she doesn't understand how her actions are affecting the classifier. She is doubting a lot of the promotion she selected on the "feature labeling" task, and she cannot find a way to tell if the classifier has improved (BUG001). She will give up quickly, blaming herself for not being able to figure it out. Tim will spend more time doing the task, especially the promoting of the words where you can undo things. But he will eventually give up, blaming the system for not providing enough feedback or an undo button (BUG011). Pat will give up, faster than Tim but not as fast as Abby, for the same reasons than Abby.

### **2.2.4 Attitude Toward Risk**

Abby and Pat's risk is the amount of time they have to commit to this task. They do not have a lot of time to do these tasks, so they will give up quickly when they do know how to achieve their goal (get feedback), nor if their input so far has been useful (BUG001, BUG002). Tim realizes that he cannot undo the labeling of examples, thus he will quickly stop tinkering on examples and tinker only on the promotion of features (BUG011). He will eventually give up after a long time because he realizes that the system cannot tell me when to stop (BUG014).

### **2.2.5 Learning: by Process vs. Tinkering**

Abby and Pat needs some direction on how to use the system (BUG006), especially when their goal changes from labeling examples to finding a feature (how to undo or how to get the classifier's performance). Tim is a tinkerer but he thinks the task is too constraint (BUG015). He cannot tinker a lot because there is a lack of undo option in task 1, and lack of feedback on the promotion in task 2 (BUG001, BUG002, BUG011).

## **2.3 Detailed Evaluation**

### **2.3.1 Step 1: User visually scans the "instance labeling" screen, searching for some feedback from the algorithm.**

Refer to Figure 8.1 in Appendix.

#### **Motivations**

At this point in the experiment, Abby, Pat, and Tim, all have labeled examples and features, so they are somewhat familiar with the interface. They are likely to know what all these buttons do. Thus, they are all likely to decide that the only thing to do is to scan the screen visually to find something they missed.

### **Information-Processing Style**

Abby and Pat will thoroughly scan the screen, while Tim will tinker with the buttons trying to find a button he hasn't used yet but he won't be able to delve a lot because the system is very constrained (BUG015).

### **Computer Self-Efficacy**

Abby has low self-efficacy, and while she is searching for a way to know if the classifier has made progress, she is likely to think she is missing something from this screen, but cannot figure out what or where it is. Pat may try the 'skip' menu, and carefully scanning the screen before giving up, like Abby. Tim might quickly try every button on the screen before realizing that none of them gives him any indication of the classifier performance. He also realizes quickly that he cannot undo previous actions, and blame the software for its lack of utility (BUG011).

### **Attitude Toward Risk**

Abby and Pat are risk-sensitive to the time they spend doing this step. They don't want to make a mistake while using the system, and spend a lot of time searching on the screen. Tim cares less about the time spent scanning the screen, he will scan the screen and try every button, even trying to type in the free text area under "Other Reasons" in the Skip menu.

### **Learning: by Process vs. Tinkering**

Abby and Pat were able to follow the process before, since the screen says "Select the correct category". They are frustrated because there is no clear instructions on the screen about getting the classifier performance (BUG006). Tim tries every single button when he scans the screen, trying to find something new that might help with his goal.

## **2.3.2 Step 2: User does not find such feedback and has to move on to the next screen, hoping to find more info on the next screen.**

Refer to Figure 8.1 in Appendix.

### **Motivations**

Abby, Pat, and Tim motivation will decrease a lot as they do not find the feature they want.

### **Information-Processing Style**

The only information they were given about this screen was during the oral tutorial with the researcher. The researcher said "In this interface, the classifier is presenting you this example. Your task is to select the label that fits best with this example. You can also skip this example if you do not know which label fits best.". However, there is literally no other information available on that screen related to the classifier's performance. Abby and Pat would be very frustrated by this screen because it never presents anything new as they go through the task. Tim might tinker with the skip button (He is likely to miss the skip button at first because the Skip button is grey instead of green.), or try to select multiple labels (he cannot), then select a label. Tim realizes too late (he didn't make sure before he clicked) that he cannot undo his labeling actions on the examples (BUG011).

### **Computer Self-Efficacy**

Abby has low self-efficacy, and since she did not find what she wanted, she is likely to think she missed something important from screens she has worked with so far. Pat will have a similar reaction than Abby. Tim realizes quickly that the system is missing some critical features, and he cannot do the scenario's goal because it is simply not possible.



### **Attitude Toward Risk**

Abby and Pat are risk-sensitive to the time they spend doing this step. They spent a lot of time searching so they might try to quickly go to the next screen. Tim will try to quickly skip through the next screen, searching for something that might have changed, like a new text displayed somewhere. He is likely to skip examples at this point to go faster.

### **Learning: by Process vs. Tinkering**

Abby and Pat were frustrated because there is no clear instructions on the screen about getting the classifier performance nor instruction indicating if the information will be displayed on the next screen (BUG006). Tim at this point might have tinkered with everything, so he is likely to stop.

### **2.3.3 Step 3: Repeat 10 times (for the ten examples)**

This screen will repeat 10 times, so that the user labels 10 examples. So we won't repeat what we said previously but we know that the frustration of repeating the search for new information will increase for all of our personas. :)

### **2.3.4 Step 4: User sees the promoting page, scans the page searching for some feedback from the algorithm.**

Refer to Figure 8.2 in Appendix.

### **Motivations**

At this point in the experiment, Abby, Pat, and Tim, all have labeled examples and features, so they are somewhat familiar with this screen. They are likely to know what all these buttons do. Thus, they are all likely to decide that the only thing to do is to scan the screen visually to find something they missed.

### **Information-Processing Style**

Abby and Pat will thoroughly scan the screen by re-reading each word, making sure it all make sense to them, maybe double-guessing themselves searching for something they missed. Meanwhile, Tim will tinker with the promotion buttons trying to figure out how it affect the system. He notices that the promoted words are always put on the top, so he is likely to not even look at the words at the very bottom of the list, because he wants to find a feature he hasn't used yet but he won't able to delve a lot because the system is very constrained (BUG015).

### **Computer Self-Efficacy**

Abby has low self-efficacy, and while she is searching for a way to know if the classifier has made progress, she is likely to think she is missing something from this screen, but cannot figure out what or where it is. Pat may carefully scan the screen before giving up, like Abby. , Abby and Pat are likely to re-read each word, making sure it all make sense to them, maybe double-guessing themselves. Tim might quickly try every button on the screen before realizing that none of them gives him any indication of the classifier performance. He blames the software for its lack of utility.

### **Attitude Toward Risk**

Abby and Pat are risk-sensitive to the time they spend doing this step. They don't want to make a mistake while using the system, and spend a lot of time searching on the screen. Because of his attitude toward risk, Tim cares less about the time spend scanning the screen, he will take time to tinker with the promotion of words and see where the words promoted end-up on the list.

### **Learning: by Process vs. Tinkering**

Abby and Pat were able to follow the process before, since the screen says “Promote words for each category”. They are frustrated because there is no clear instructions on the screen about getting the classifier performance (BUG006). They are also frustrated because sometimes words can be promoted for two categories (e.g. “brew” or “beans”) and they don’t know what they should do (i.e., promote in both category or none). Tim will tinker with the promotion of words and see where the words promoted end-up on the list. He is likely to realize that the promoted words always end-up at the top of the list and this is how the classifier interprete usefulness.

### **2.3.5 Step 5: User does not find such feedback and has to move on to the next screen, hoping the end is nigh.**

Refer to Figure 8.2 in Appendix.

### **Motivations**

Abby, Pat, and Tim motivation will decrease a lot as they do not find the feature they want.

### **Information-Processing Style**

The only information they were given about this screen was during the tutorial with the researcher. The researcher said “This page provides you a list of words for each label. The classifier thinks that these words are useful to identify a particular label. Your task is to promote words, if you think they are useful to recognize a label. Once you are done, you can select “Continue””. However, there is literally no other information available on that screen related to the classifier’s performance. Abby and Pat would be very frustrated by this screen after spending so much time reading and re-reading the same words. He realizes quickly that this screen does not offer any information about the performance of the classifier.

### **Computer Self-Efficacy**

Abby has low self-efficacy, and since she did not find what she wanted, she is likely to think she missed something important from screens she has worked with so far. Pat will have a similar reaction than Abby. Tim realizes quickly that the system is missing some critical features, and he cannot do the scenario’s goal because it is simply not possible.

### **Attitude Toward Risk**

Abby and Pat are risk-sensitive to the time they spend doing this step. They spent a lot of time searching so they might try to quickly go to the next screen or just give up. Tim will try to quickly skip through the next screen, searching for something that might have changed, like a new text displayed somewhere. He is likely to not even re-read the list to go to the next screen before ultimately giving up on his goal.

### **Learning: by Process vs. Tinkering**

Abby and Pat were frustrated because there is no clear instructions on the screen about getting the classifier performance nor instruction indicating if the information will be displayed on the next screen (BUG006). Tim at this point might have tinkered with everything, so he is likely to stop.

### **2.3.6 Step 6 - Repeat ad nauseam**

The user vomits on the researcher because reading endless, purposeless, and repetitive list of words is nauseating. :) (BUG014)

# Chapter 3

## User Observation

### 3.1 Research questions

Our main goal is to discover “What are the difficulties encountered by a user?”

### 3.2 Process to answer RQ

To answer this research question, we did a heuristic evaluation using GenderMag and Gestalt heuristics. We found that the main problem in our system is the feedback to the user. In addition, there are some small usability issues related to effective invisibility of options. To confirm these issues, we performed a user observation to see if the problems we foresaw actually happened.

### 3.3 User Observation

#### 3.3.1 Scenario

Abby wants to know if she is being consistent in her labeling or if she has inadvertently introduced more uncertainty in the learning agent.

#### 3.3.2 Observations

The people in the space

1. Who are they, what are they like?
  - She is student majoring in MBA
  - She was nervous before she started to test
2. What are they doing?
  - She was writing her assignments before starting the experiment.
3. How are they doing it?
  - Using her computer at the library.
4. What do their emotions, purposes, reactions seem to be?
  - First, when I was reading the script (attached), she looked very focused.
  - After looking at the first example, she looked confused because the first example was a garbage example (e.g., “etc.”). The dataset is not perfectly cleaned (BUG005).
  - When she encountered an example that she couldn’t decide, she turned to us for help (BUG006). Due to the experimental setting, we could not give her any instructions about the “right” answer to label the example, she looked frustrated and decided to guess the label that best fit her emotion and moved on to the next example. When she got to the promoting words’ page, she was very upset about not knowing which button to click and what the consequence of promoting a word was for the classifier (BUG002).

- At the first iteration of the promotion, she decided to not click on any words and continued to the next page (BUG006).
  - In the first round of labeling, she noticed the skip button, and decided to use it when she was not certain about the category. There was one time she typed “I don’t understand the question” in the feedback box under the skip menu (BUG005). There was more examples she couldn’t understand, so she started to shake her leg, and looked like it is starting to be a little annoying. And at one point, she yawned.
  - After second round of examples, she became impatient and she started to use one hand to support her head. During the second promotion, although she decided to click on the green button, she was very hesitant to do so.
  - As the time goes by, she started to be even more impatient, and was eager to finish the questions.
5. What problems do they encounter with their activities?
- She was afraid that she chose the wrong labels (BUG006).
  - She didn’t understand the examples because the dataset can be messy. (e.g., out of context sentences, or too short to be understood in the context of the labeling task.) (BUG005)
  - She didn’t discover that she can “skip the example” until later on. (BUG008)
  - She thought this experiment was boring and wasted her time.

### **The objects (technological and otherwise) in the space and with the people**

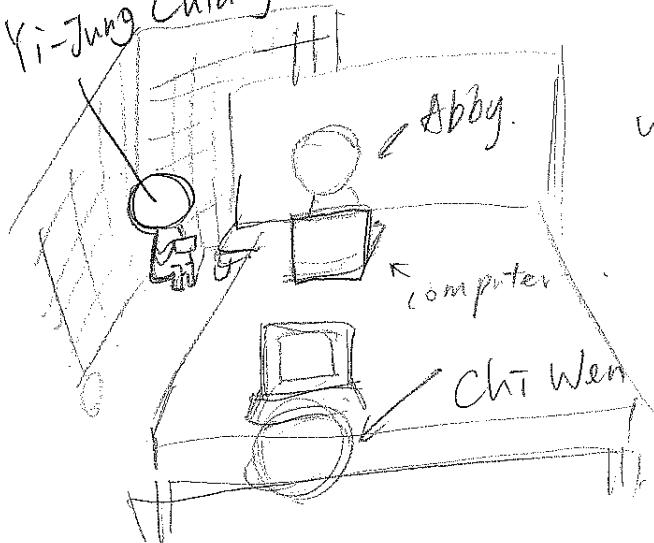
1. What are the functional elements of the objects?
  - Long couch, large table, two computers (one for the researchers, one for the user), pens and papers.
2. What are the decor elements?
  - Partition, carpeted floor, window.
3. Which objects do people look for (perhaps to somehow interact with)?
  - The user only looked at her computer and toward the researchers, to make eye contact when asking for help.
4. Which objects do people bring with them that matter to the activities they are trying to do?
  - The user didn’t bring their own computer but used one of the researcher’s computer to use the system, because the system was available only on a researcher’s computer. It is not available online.

### **Environment: spaces, architecture, lighting etc**

1. What is the layout?
  - A small cubicle in the library. Carpeted floor, long couch, large table.
2. What is the environment like?
  - We took place in the library, which is full of students. Fortunately, we found a small space with partition around the area.
3. How does it influence the activities people engage in?
  - Just few people walk around and most people talk to each other with low volume so the activities people can easily focus on our project.
4. How does the environment support the objects above?
  - She is more focused. She won’t be distracted by other students.

## 3.4 Raw Data

Yi-Jung Chiang (video recording)



5/4

2:00 PM

Second floor  
Library

Abby

- MBA student

-(Nervous)

2:05 PM Focusing to introduction (about Charlie), but she looked confused

2:10 PM ① start to test

② When she saw the website, she didn't know how to do and she didn't know example of meaning (e.g. "etc.") she was still confused and wanted us to help her.

2:17 PM

she noticed that she can select "skip the example" and it's first time she clicked this button.

2:19 PM

Entered "promote word" part but she directly clicked "continue" without promoting words.

2:22 PM

2:24 PM she felt impatient, shake leg.

2:29 PM

Entered "promote word" section again but she promoted some words this time. However, she hesitated them.

2:33 PM she wanted to finish

## 3.5 Conclusion

With the detailed observations, point out the places that provide Results/Insights and say what they are ("I"):

### 3.5.1 What are the answers to your research questions?

The difficulties encountered by the user seems to be related to process - what should I do? How should I do it? - and feedback - What are the consequences of doing it? These questions are left unanswered by our system and create a lot of frustration for the user.

### 3.5.2 What other insights did you get from this that are relevant to your "patient"?

Natural datasets can be messy, and this is also a source of confusion for the user. There is no clear action available for the user to deal with an example that should be clearly ignored by the classifier.



# Chapter 4

## Other Heuristics

### 4.1 Gestalt Principles

- Proximity : (BUG017) When promoting the words, in case of reminding the user of the category, the category named repeatedly every 20 words. It causes a confusion of identify it as two different group.
- Continuity: (BUG016) The list of words ends every 10 words to add the label of the category. It makes the table look like it is finished, and the user does not realize they have to scroll down to continue the lists.

### 4.2 Trap & Tenets

- Effectively Invisible Element (T#2): (BUG008) The skip button is located at the bottom of the page, away from the 'Next' button and the labels button, therefore it is effectively invisible to the user.
- Memory Challenge (T#8): (BUG003) When promoting words, the user might forget their previous answers and want to double check them to stay consistent. The system doesn't provide any function for user to do so.
- Feedback Failure (T#9): (BUG0013) If the back-end system crashes, the Wait screen will stay on forever. The user is not informed that the back-end system crashed.
- Unnecessary Step (T#14): (BUG012) There are two buttons hidden inside the Skip menu. Skipping takes two steps, where 'Continuing' to the next example takes one step.
- Information Overload (T#16): (BUG009) After doing several sets of feature selection, the top of the feature lists will be cluttered with words, which is redundant and tedious for user to read and review.
- Irreversible Action (T#18): (BUG004) The system does not provide an undo button when labeling examples or review previous labels.

# Chapter 5

## Concept

### 5.1 Bug Fixes

For a complete list of bugs, refer to Table 8.1 in Appendix.

**BUG001 No feedback on the classifier accuracy** : From any main screens (#4,5,6, 7, 9 or 10), we added a frozen menu bar. The menu bar presents an icon (a graph). By clicking on the icon, a small window appears ("Summary" on screen #2) and presents a summary/XAI explanation of the current state of the classifier.

**BUG002 No feedback about the consequences of the user's actions onto the classifier** : In addition to the fix from BUG001, we added more feedback related to the promoted words. When the user click on a promoted word (screens #5,6,9,10), a window pops-up and presents some feature specific explanation of the current and past state of the feature.

**BUG003 Memory Challenge T#8 UI does not provide a way to view previously labeled example**: Screen #7 presents a list of all labeled examples. Screen is accessible at any point in the activity, through the menu bar (the list icon).

**BUG004 Irreversible Action T#18/Unsafe system/ UI does not provide a way to change labels on previously labeled examples**: On screen #7 and 8, users can edit the labels, they can also "ignore" an example, which essentially delete the example from the classifier's dataset. During one batch of example (Screen #4), the user can use the 'Back' button to modify examples from the current batch.

**BUG005 Some examples are garbage data** : The dataset is not perfectly clean and there was no way to deal with it. The user can skip an example (screen #4) and later on decide to ignore it completely (screen #7). Ignoring an example will effectively remove it from the dataset visible to the classifier.

**BUG006 No instructions about how the system works** : On screen #1, we added a "FAQ" screen accessible through the '?' icon on the top menu. The "FAQ" condenses the main working point of the system (Whats and How-tos), as well as a search bar for quick access to relevant information.

**BUG007 Users does not know whether they can cancel words which they already chose**: The feature important activity (previously "list of words" screen) has a separate list of "promoted" words, with a little cross next to each word for easy deleting<sup>1</sup>.

**BUG008 Effectively invisible T#2/Gestalt proximity violated - Skip button** : On screen #4, moved the 'Skip' button next to the 'Next' button.

**BUG009 Information Overload T#16: The promoted words clutter the top of the list**: The feature importance screen (Fig. 8.2) is completed re-designed. Now, the important of words for each class is presented in a word cloud where the importance of a feature corresponds to its size in the cloud. Once a word is selected for promotion, it is moved to

---

<sup>1</sup>The cross isn't on the scan but is on the poster version of the concepts

the adjacent list, and does not clutter the cloud. The cloud does not grow beyond a certain word limit, not like the list which could grow ad vitam eternam.

**BUG011 UI does not provide undo/redo button** : On screen #4 , added a 'Back' button for user to go back to the previous examples in the same batch. Also, a list icon is added on the top-right of the screen, which can take the user to screen #7. On screen #7, a list of previous examples (beyond the current batch) is added for the user to check/change their answers.

**BUG012 Unnecessary steps T#14 - Menu inside skip button is not useful** : On screen #4, replaced the skip menu with single 'Skip' button.

**BUG014 No way to actually finish a session** : On screen #4,5,6,9,10, the "finish button" is added on the top-right of the screen. The summary page is automatically displayed after choosing to finish the session. A report can be saved from the summary page.

**BUG015 System is very constrained (unable to delve a lot)**: Screens #1,2,3 and 7 allow much more options in how to interact and learn from the system. Tinkering is now possible.

**BUG016 Continuity - The list of words ends every 20 words to add the label of the category. It makes the table look like it is finished, and the user does not realize they have to scroll down to continue the lists.:** On screens #5, #6, #9, and #10, we changed the list of words to a word cloud display. The user clicks on a word in the word cloud to promote it. The promoted words appear on a separated list on the right. The list of words is not interrupted every 20 words and provided with a scroll bar on the side.

**BUG017 Similarity: When promoting the words, in case of reminding the user of the category, the category named repeatedly every 20 words. It causes a confusion of identify it as two different group.:** The feature importance screen (Fig. 8.2) is completed re-designed. On screens #5, #6, #9, and #10 the importance of words for each class is presented in a word cloud where the importance of a feature corresponds to its size in the cloud. Each class is on a separated panel.

# Chapter 6

## Scenario 1

Scenario 1 is "Abby wants to check and/or modify a label she has entered during a previous instance labeling iteration"

Originally, there wasn't any function for Abby to check her previous labels. We added a menu bar at the top of the system (see on Figures 8.7, 8.4, 8.10). This menu bar follows the principles of consistency for being identical in every screen and does the same thing everywhere. Also, it applies the Gestalt principles of Proximity and Similarity, to express the idea of a different set of options of the system, which are different from the main task. Using feedback from Dr Burnett during the final presentation, we added labels under each icon, to indicate Affordance.

This bar contains an icon for performance<sup>1</sup> (labeled "Results"), an icon for help (labeled "FAQ"), an icon to access the list of previously labeled instances (labeled "Previous Examples"), and an icon to quit the program (labeled "Exit"). The icon for performance takes the user to the Summary screen (Figure 8.15, BUG001, BUG002), the help icon will make the FAQ pop-up appear (Figure 8.5, BUG006), the list icon will bring the user to a page where user can edit or remove their previous labels (Figure 8.10, BUG003, BUG004, BUG005, BUG015), and finally, when the user clicks on the exit icon, it will pop-up an alert for user to choose whether they want to exit the session or not (Figure 8.6, BUG014).

Additionally, it was mentioned during the final presentation that a "Save" button would be easier than having to go to the Summary page. We chose not to add a "Save" icon to the menu bar to keep the UI as simple and uncluttered as possible. It would add visual complexity (one more label, one more icon), and be redundant because you can save from the Summary page and the Exit button. Note that every labeling action is saved automatically. While the user might not know that explicitly, they are do inquire about this particularity of the system, it will be mentioned in the FAQ under "How do I save my work?" and again under "How do I exit?".

### 6.1 List Page

In this scenario, Abby wants to check or modify a previous label. To do so, she clicks on the list icon on the top-right menu icon "List". It brings her to the List page (Figure 8.10). By providing a full list of the examples and their chosen label, we remove the burden of memory challenge from the user (BUG003). By providing the option of editing a label (Figure 8.14), we remove the Irreversible Action problem (BUG004). By adding an option to ignore completely an example (essentially deleting it from the dataset), we remove garbage data (BUG005). Both options (edit and ignore) are located close to each other, in the same box, respecting Gestalt principles of Proximity to signal that they apply to only one specific example.

Overall, this gives more flexibility and possibilities of branching out for a Tinkerer. Thus, we somewhat resolve some of the constraint problem (BUG015). We will address more way to alleviate constraint below.

---

<sup>1</sup>The same icon is used in Screen 8.7 next to the promoted feature. We did not add the label on that screen because of space, but we hope that by using consistency in imagery, and habituation, the user will learn that this icon means "get more info about the performance at that point and in this context" even if the context is different.

# Chapter 7

## Scenario 2

Scenario 2 is "Abby wants to know if her previous feedback to the classifier has had any impact on its performance.

In the previous version of the software, there wasn't anyway for the user to obtain the classifier's current or past performance (BUG001 and BUG002). In order to solve these feedback problems, we implemented two ways of getting an idea of what and how the classifier is doing.

### 7.1 Summary Page

We implemented a menu bar at the top of every screen (excluding pop-ups, see Figures 8.7, 8.4, 8.10). In the previous chapter, we described the principles we used to create this menu. In this section, we will talk about the performance icon and how it solves BUG001 and BUG002.

The icon for performance is a little graph icon for visibility and affordance. It takes the user to the Summary screen (Figure 8.15). The Summary screen will present some detailed information about the classifier's performance, and feedback about the user's input so far. This information will be interpretable by a human who is not a machine learning expert. Unfortunately, the precise content of this page is out of the scope of this project.

This page addresses BUG001 and BUG002 by providing contents about the overall feedback to the user about the classifier performance.

### 7.2 Feature Importance Explanation Page

Abby can also learn more about the effect of promoting features via the feature importance explanation screen (Figure 8.9). During the Feature Importance labeling task (Figure 8.7), Abby can click on a feature in the word cloud to "promote" it. The feature promoted appears in the list on the left. Once a word has been promoted and placed in the list, Abby can click on it to obtain more information. Figure 8.9 shows the pop-up appearing after Abby clicked on "Brew". In order to fix the feedback problems we were having (BUG002), this page will provide humanly interpretable explanation of the effect of promoting this feature, as well as how important this feature is to the classifier and "why". This information will be interpretable by a human who is not a machine learning expert. Unfortunately, the precise content of this page is out of the scope of this project.

## Chapter 8

### Scenario 3

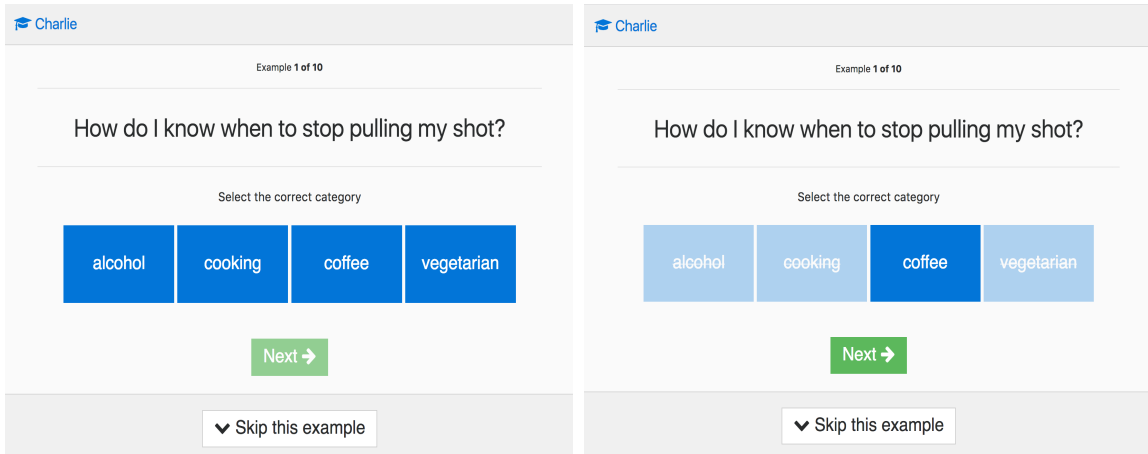
Scenario 3 is: "Abby wants to save her work and stop working on this."

In the previous version of the software, users could not "officially" finish a session. They could stop working and their session was saved but they didn't know it since the system was setup as an experimental tool with no concern for proletarian needs such as "stop working" ;). We added an "exit" button in menu bar on the top right corner of the screen to resolve BUG014 (See previous chapters about what principles apply to menu bar).

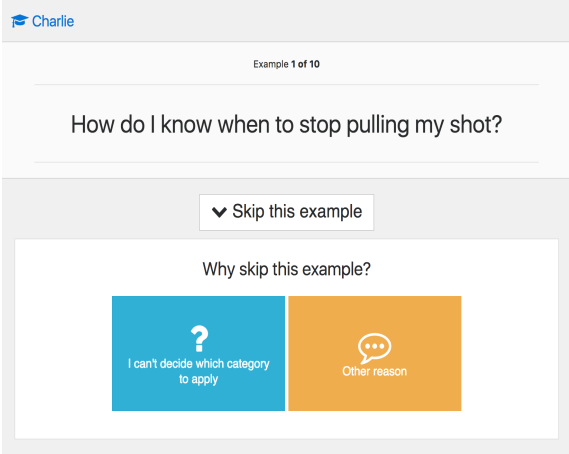
Now, in order to exit the system, the user can click on the "exit" button. For Safety, an "Alert" message interrupts the activity to confirm the action (See on Figure [8.6](#), [8.8](#), [8.13](#)). The alert prompts the user that the session will automatically be saved. The screens to recover a saved session are out of the scope of this project.

Another way to exit the system is offered via the Summary page (See on Figure [8.15](#)). An "Exit" button is afforded next to the "Save" button. On again, for Safety, an "Alert" message interrupts the activity to confirm the action (Figure [8.16](#)).

# Appendix

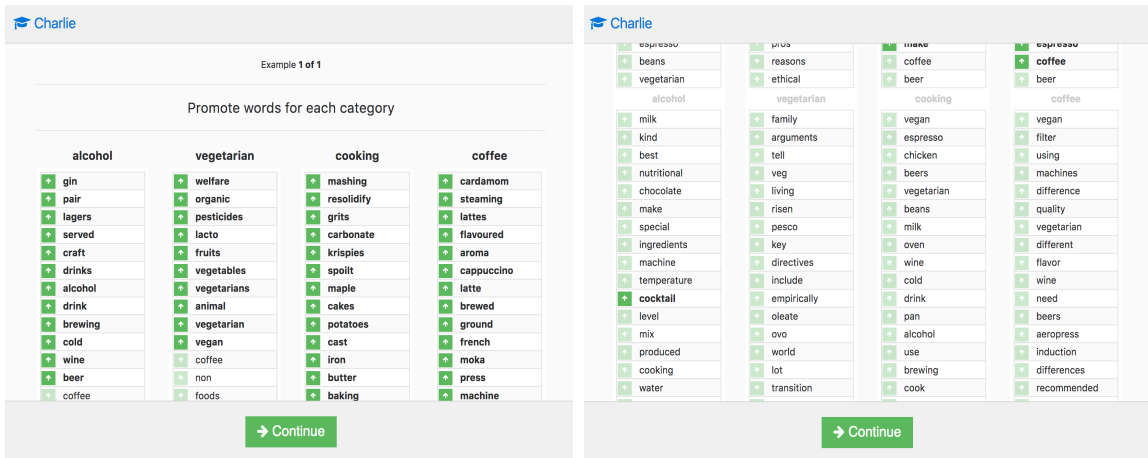


(a) Instance Labeling UI (b) Instance Labeling UI when label is selected



(c) Skip menu

Figure 8.1: Instance Labeling UI



(a) Feature Labeling UI (b) Feature Labeling UI, when scrolled down

Figure 8.2: Feature Labeling UI



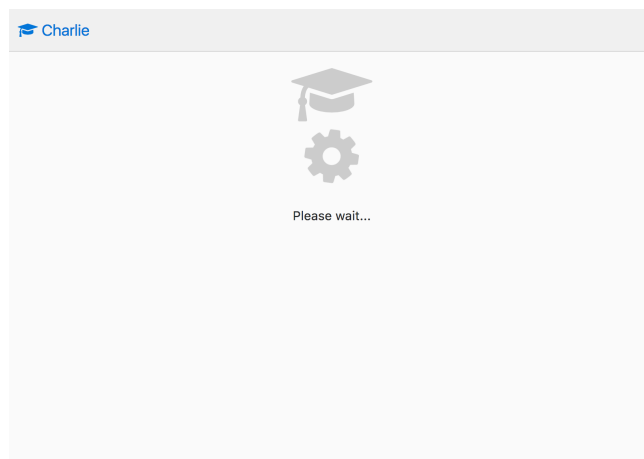


Figure 8.3: Waiting Screen

BugID	Description
BUG001	UI does not provide feedback on the classifier accuracy
BUG002	UI does not provide feedback about the consequences of the user's actions onto the classifier
BUG003	Memory Challenge T#8/ UI does not provide a way to view previously labeled example
BUG004	Irreversible Action T#18/Unsafe system/ UI does not provide a way to change labels on previously labeled examples
BUG005	Some examples are garbage data (the dataset is not perfectly clean) and there is no way to deal with it
BUG006	UI does not provide detailed instructions about how the system works
BUG007	Users does not know whether they can cancel words which they already chose
BUG008	Effectively invisible T#2/ Gestalt proximity violated - Skip button
BUG009	Information Overload T#16: The promoted words clutter the top of the list
BUG011	UI does not provide undo/redo button
BUG012	Unnecessary steps T#14 - Menu inside skip button is not useful
BUG013	Feedback Failure T#9 - If the backend crashed, the Wait screens stays on forever
BUG014	No way to actually finish a session
BUG015	system is very constrained (unable to delve a lot)
BUG016	Continuity: The list of words looks finished, and the user does not realize they have to scroll down to continue the lists.
BUG017	Similarity: When promoting the words, in case of reminding the user of the category, the category named repeatedly every 20 words.

Table 8.1: List of Bugs

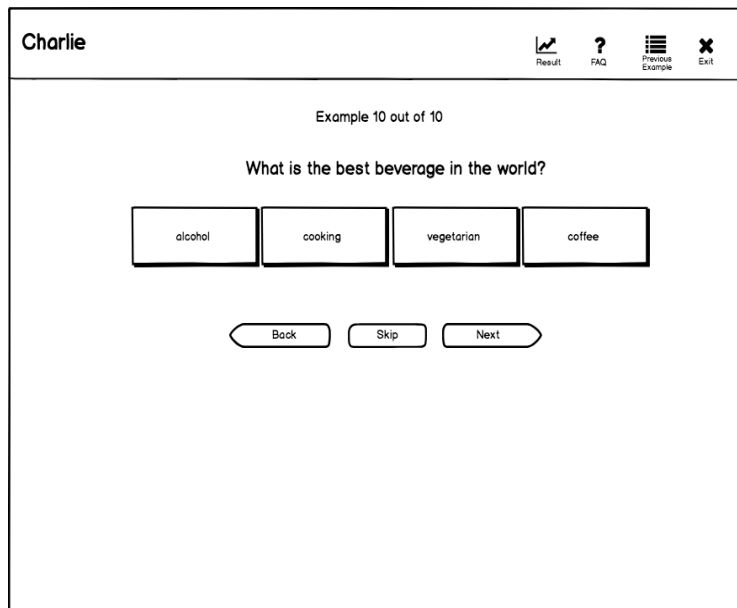


Figure 8.4: Screen for labeling an example

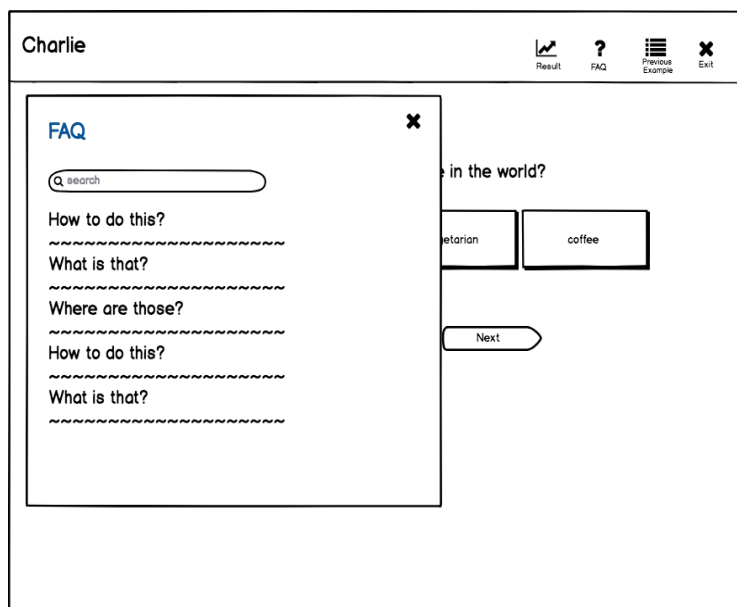


Figure 8.5: FAQ screen

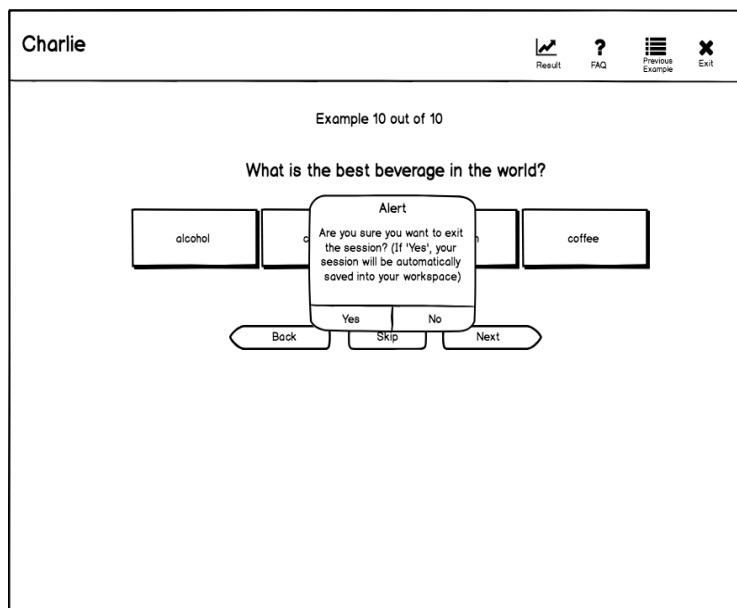


Figure 8.6: Alert exit

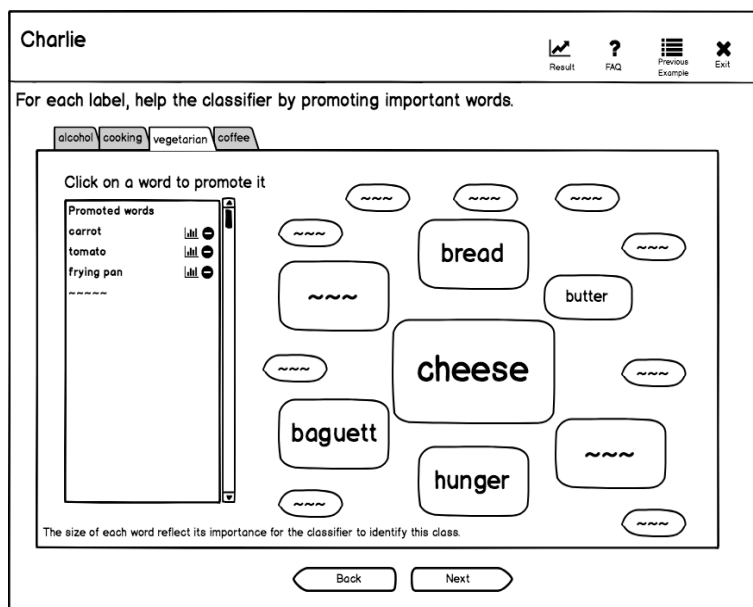


Figure 8.7: Screen for labeling a feature

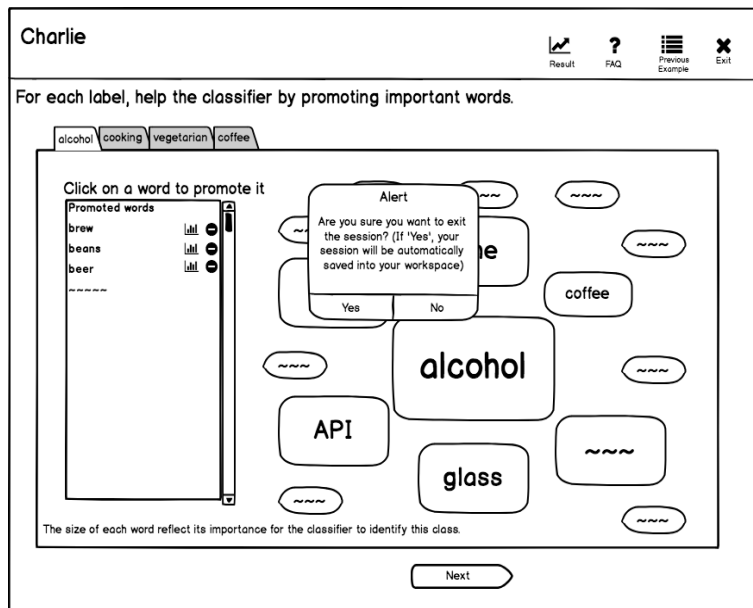


Figure 8.8: Alert exit

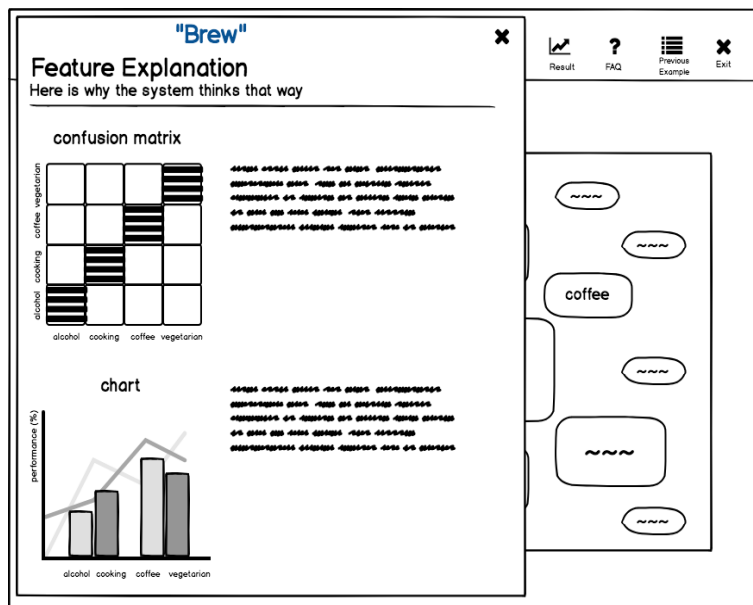






Figure 8.9: Explanation of a feature importance for the classifier

[illegible]

Charlie

 Result
  FAQ
  Go Back
  Exit

Example #	Example	Label Selected	Options
1	An example about something	alcohol	<a href="#">Edit label</a>   <a href="#">Ignore Example</a>
2	An example about nothing	vegetarian	Edit Label   Ignore Example
3	An example about something else	alcohol	Edit Label   Ignore Example
4	.....	cooking	Edit Label   Ignore Example

Alert

Are you sure you want the classifier to ignore this example?

YesNo

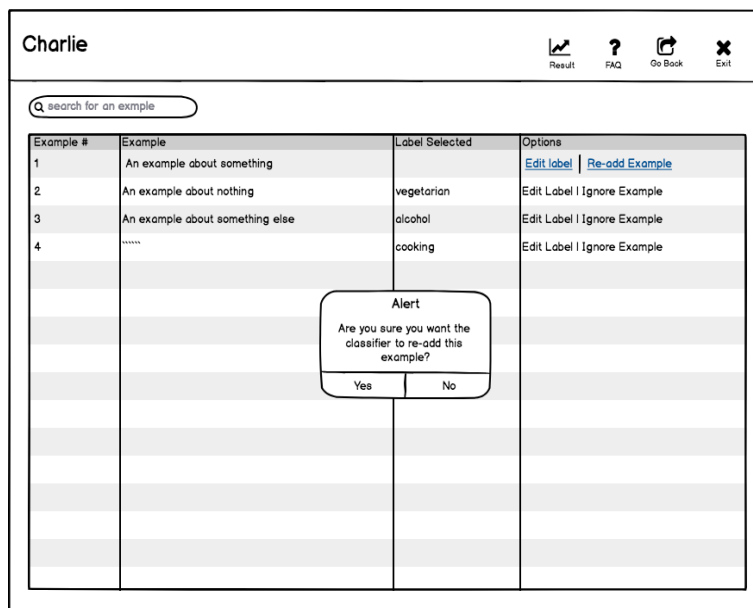


Figure 8.12: Screen to un-ignore (or re-add) an example

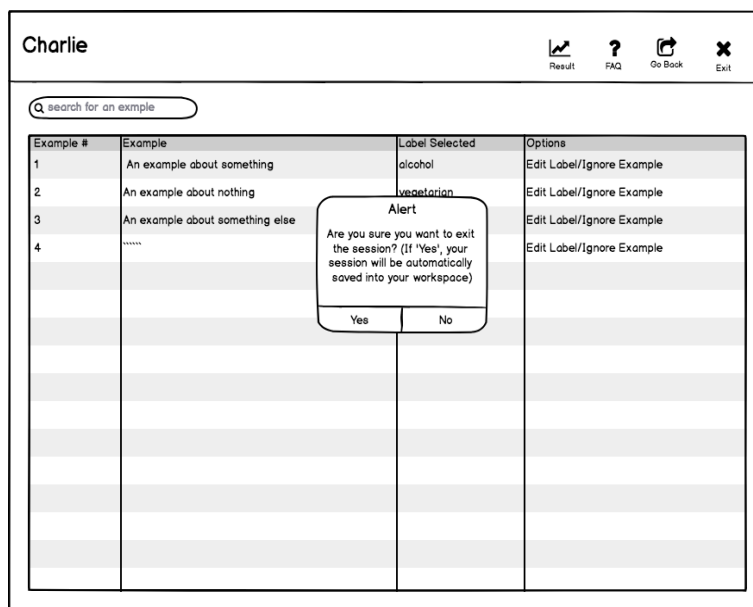






Figure 8.13: Alert to exit

Charlie

 Result
  FAQ
  Go Back
  Exit

Example #	Example	Label Selected	Options
1	An example about something	alcohol	<a href="#">Edit label</a>   <a href="#">Ignore Example</a>
2	An example about nothing	vegetarian	Edit Label   Ignore Example
3	An example about something else	alcohol	Edit Label   Ignore Example
4	.....	cooking	Edit Label   Ignore Example

Select a Label

Pick a Label

alcohol  
cooking  
vegetarian  
coffee  
Skip

Save

[illegible]



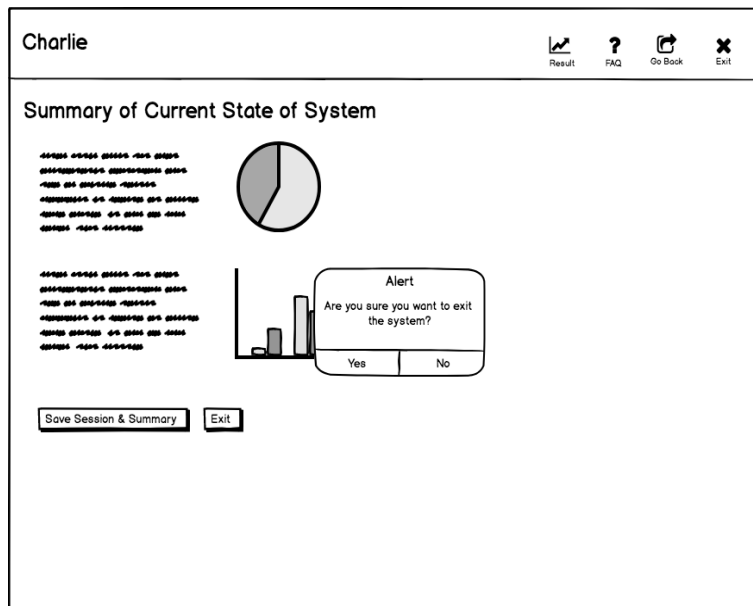


Figure 8.16: Summary screen with Exit alert

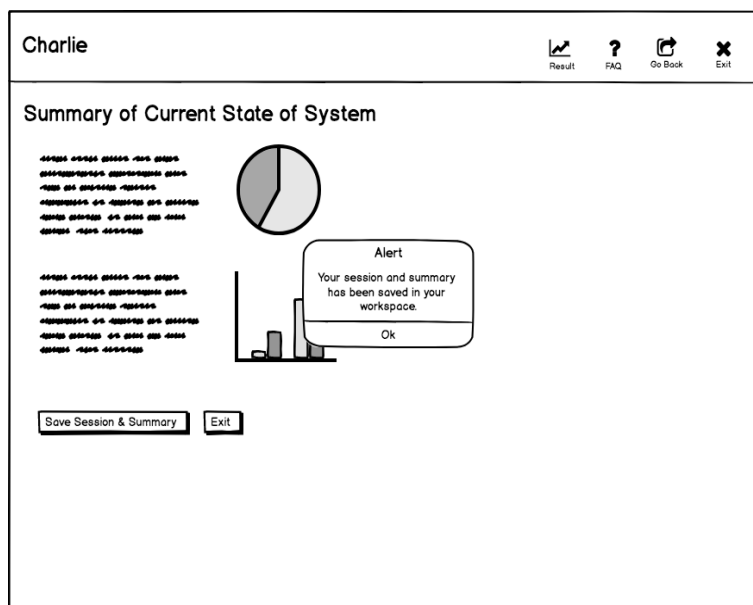


Figure 8.17: Saved Session

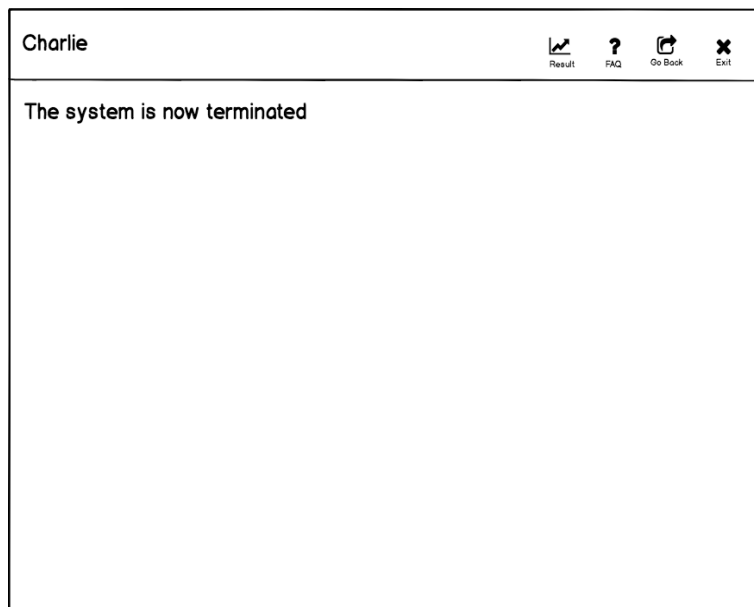


Figure 8.18: System is terminated