Be Inaccurate but Don't Be Indecisive: How Error Distribution Can Affect User Experience

Rafael R. Padovani,[1] **Lucas N. Ferreira,**[2] **Levi H. S. Lelis**[1]
[1] Departamento de Informática, Universidade Federal de Viçosa, Brazil
[2] Department of Computational Media, University of California, Santa Cruz, USA
rafael.rpadovani@gmail.com, lferreira@ucsc.edu, levi.lelis@ufv.br

Abstract

System accuracy is a crucial factor influencing user experience in intelligent interactive systems. Although accuracy is known to be important, little is known about the role of the system's error distribution in user experience. In this paper we study, in the context of background music selection for tabletop games, how the error distribution of an intelligent system affects the user's perceived experience. In particular, we show that supervised learning algorithms that solely optimize for prediction accuracy can make the system "indecisive". That is, it can make the system's errors sparsely distributed throughout the game session. We hypothesize that sparsely distributed errors can harm the users' perceived experience and it is preferable to use a model that is somewhat inaccurate but decisive, than a model that is accurate but often indecisive. In order to test our hypothesis we introduce an ensemble approach with a restrictive voting rule that instead of erring sparsely through time, it errs consistently for a period of time. A user study in which people watched videos of Dungeons and Dragons sessions supports our hypothesis.

Introduction

System accuracy is a crucial factor influencing user experience in interactive and autonomous systems. For example, de Vries, Midden, and Bouwhuis (2003) and Desai et al. (2012) have linked system accuracy to trust in autonomous systems, and according to Lee and See (2004), the wrong level of trust can lead to system misuse or disuse. While it is known that system accuracy can affect user experience in terms of trust, little is known about how the distribution of system errors affects user experience. For example, if system errors are inevitable, should one prefer a system whose errors are distributed uniformly during the system's execution or a system whose errors occur in a period of time and is accurate for the rest of its execution?

In this paper we study how the error distribution of an intelligent system affects the user's perceived experience. Our work is motived by the application domain of automatic background music selection for tabletop games. Padovani, Ferreira, and Lelis (2017) introduced Bardo, a system that uses supervised learning to identify through the players' speech the emotion in the story being told in sessions of Dungeons and Dragons (D&D),[1] a storytelling-based tabletop game. Bardo chooses a background song to be played according to the identified emotion. The domain of background music selection for tabletop games is interesting because it allows one to evaluate intelligent systems interacting with humans in a scenario in which system errors are inevitable and directly affect the users' perceived experience.

In this paper we show that supervised learning algorithms that solely optimize for prediction accuracy can make the selection system "indecisive". That is, the system's errors can be sparsely distributed throughout the game session. We call these sparsely distributed errors *short misclassifications*. We hypothesize that short misclassifications can harm the user's perceived experience and that it is preferable to use classification models that are somewhat inaccurate but decisive, than models that are accurate but often indecisive.

In order to test our hypothesis we introduce an ensemble classifier (Dietterich 2000), which we name EC, to minimize short misclassifications. EC changes the background music only if all models composing its ensemble agree with the change. As a result of its voting rule, instead of erring sparsely throughout a game session, EC tends to err consistently for a period of time. In addition to showing that EC is expected to reduce the number of short misclassifications, we derive the conditions in which EC is expected to be more accurate than a classifier c. Namely, we show that if the emotion of the story does not change "too often" and c is "not too accurate", then EC will be more accurate than c.

Empirical results on *Call of the Wild* (CotW), a set of game sessions of D&D available on Youtube,[2] show that EC is able to dramatically reduce short misclassifications while being slightly more accurate than the models in the ensemble. We test our hypothesis that short misclassifications can harm the user's perceived experience with a user study. In our study, people watched video excerpts of CotW with the background music selected by EC and by Bardo's original model. In order to test our hypothesis, we select excerpts in which Bardo's model is either as accurate or more accurate than EC and performs more short misclassifications than EC, thus giving Bardo's model an advantage in terms of accuracy. We approximate the user's perceived experience with

[1] http://dnd.wizards.com
[2] https://www.youtube.com/watch?v=tZWU5iPjQpI

the participants reported preferences for the music selected by the two approaches. The results of the study support our hypothesis that sparsely distributed errors can negatively affect the user's perceived experience and that it might be preferred to be somewhat inaccurate and decisive than accurate and indecisive in the context of background music selection.

Although we evaluate our system with an ensemble of classifiers in the domain of background music selection, we expect our results to generalize to other classifiers and application domains. That is, in principle, any classifier that solely accounts for prediction accuracy can suffer from indecisiveness. Moreover, the negative effects of indecisiveness might arise in other application domains such as recommendation systems for financial investments (Chou et al. 1996; Seo, Giampapa, and Sycara 2004).

Related Work

The relation between system accuracy and user experience in terms of trust has been extensively studied, see the work of Yang et al. (2017) for a recent example. Most works on accuracy and trust involve one manipulating the error distribution of an autonomous system and measuring the user's trust on the system. Sanchez (2006) controlled system errors to occur either in the first or the second half of a simulated task. They found that the users' trust was significantly lower if the errors occurred on the second half of the simulation. Desai et al. (2012) observed that users tend to switch an autonomous system to manual mode more often if the system errors occur in the middle of a task. In addition to studying the impact of system errors in user perceived experience, we introduce a learning model that shifts the prediction error distribution of a music selection system. Also, we measure the impact of the error distribution on how the users perceive the background music, and not the user's trust on the system.

An ensemble of diverse classifiers that perform slightly better than random guessing is known to result in accurate models (Hansen and Salamon 1990; Schapire 1990). Several algorithms were developed to create such ensembles. In Bagging, one trains several classifiers with a different sampling of the training data (Breiman 1996). AdaBoost also manipulates the training data by applying a weight to the training error of each training instance (Freund and Schapire 1997). Another way to create a set of potentially diverse classifiers is by training models on different subsets of the instances' features (Cherkauer 1996). We use an approach similar to Cherkauer's as we train a set of classifiers with and without a feature selection procedure to compose EC's ensemble. Nonetheless, one could use any of the previous approaches to train a set of diverse classifiers to compose EC's ensemble. In contrast with other ensemble methods which primarily try to improve prediction accuracy, EC is designed to alter the error distribution of its base classifiers.

The problem of identifying which song to play in sessions of a tabletop game has a temporal structure that could be better captured by other models such as Hidden Markov Models and Long-Short Term Memory networks (Hochreiter and Schmidhuber 1997). Moreover, aiming at having an accurate and decisive system, instead of training an ensemble of classifiers, one could train a model that directly minimizes short misclassifications and maximizes prediction accuracy. Although we could have used in our experiments any of the approaches mentioned above, we note that the contribution of evaluating how the system's error distribution affects user perceived experience is algorithm agnostic. This is because system indecisiveness can potentially occur with the use of any classifier that is solely concerned with prediction accuracy. The use of an ensemble of classifiers is a solution to the short misclassification problem that allowed us to test our hypothesis that sparsely distributed errors can harm how the user perceives their experience in the context of background music selection. We expect future works to develop other solutions to the system indecisiveness problem.

Bardo

Padovani, Ferreira, and Lelis (2017) introduced Bardo, a system that automatically selects background music for tabletop games. Bardo uses a speech recognition (SR) system to translate into text what players say during a game session. Since Bardo was originally tested with Youtube videos, Padovani et al. employed Youtube's SR system, which is normally used to generate subtitles, to convert speech into sentences. We follow the same approach in this paper and what we refer as a sentence we mean a subtitle generated by the SR system. A Naive Bayes (NB) approach was then used in Bardo to classify each sentence being produced by the SR system into one of the four "story emotions": Happy (H), Calm (C), Agitated (A), and Suspenseful (S). Bardo selects a song from a library of songs that corresponds to the current classified emotion (Bardo requires the songs to be labeled according to the four emotions). The selected song is then played as background music in the game session. Bardo operates in real time and switches the background music whenever NB detects an emotion transition in the story.

Padovani et al. used a sliding window approach in which, instead of providing only the last sentence produced by the SR system to the classifier, Bardo provides the last z sentences produced by the SR system. Also, as argued by Padovani et al., the sentences provided by the SR system are often noisy as the system captures the speeches of all players simultaneously. For that reason, Padovani et al. did not extract the structure of the sentences and used the sentences as a bag of words for classification. We take the same approach in this paper. Whenever referring to a bag of words, we are referring to the last z sentences returned by the SR system.

We use Padovani et al.'s dataset, which includes 9 episodes of CotW. The dataset contains 5,892 sentences and 45,247 words, resulting in 4 hours, 39 minutes, and 24 seconds of gameplay. There are 2,005 Agitated, 2,493 Suspenseful, 38 Happy, and 1,356 Calm sentences in the dataset.

Sparsely Distributed Errors

Bardo was originally evaluated in terms of prediction accuracy. However, system accuracy does not provide information about the system's error distribution. The distribution of errors is important in the domain of music selection because even an accurate but imperfect system could harm user experience if it switches the background music too often.

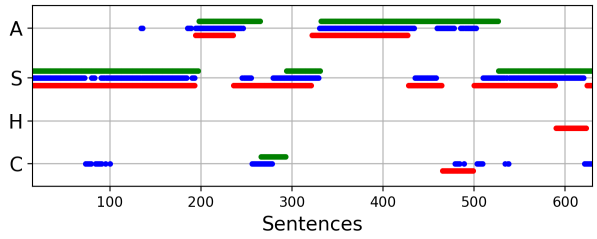

Figure 1: Comparison between the true labels (red lines), NS's classifications (blue lines), and EC's classifications (green lines) for an excerpt of episode 6. The y-axis shows the four different story emotions and the x-axis the sentences in the order of appearance in the episode.

Figure 1 shows the actual emotions and the classifications of two models in an excerpt of episode 6 of CotW. The x-axis shows the sentences of the episode ordered by appearance, and the y-axis the emotions. The red lines show the actual labels of the sentences, the blue lines show the classifications of Bardo's original model, called NS, and the green lines show the classification of EC. As an example of how to read Figure 1, both NS and EC are accurate in the beginning of the episode as their classification match the actual emotion.

Although NS is more accurate than EC in episode 6 (NS has an accuracy of 69% and EC of 59%), we hypothesize that NS is more likely to harm the players' perceived experience than EC. This is because EC is consistent in the sense that its errors are not sparsely distributed in the excerpt. For example, NS quickly switches between the Agitated and Calm emotions around sentence 500. We quantify these switches in terms of short misclassifications (SM),

$$SM(C, W) = |T(C, W) - T(W)|\,.$$

Here, $C(w)$ is a classifier that receives as input a bag of words w and returns for w a class in Padovani et al.'s emotion model. Also, $W = \{w_1, w_2, \cdots, w_m\}$ is an ordered collection of bag of words, whose order is defined by the order of appearance of the m sentences of an episode. That is, w_i is the bag of words of the i-th sentence appearing in the episode. We refer to the subscript i as the time step in which the bag of words appears in the episode. $T(C, W)$ is the number of times C classified two adjacent sentences in W with different labels, formally defined as,

$$T(C, W) = |\{w_i | w_i \in W \wedge C(w_i) \neq C(w_{i+1})\}|\,.$$

$T(W)$ is the number of actual emotion transitions in W,

$$T(W) = |\{w_i | w_i \in W \wedge L(w_i) \neq L(w_{i+1})\}|\,.$$

Here, $L(w)$ is the true label of sentence w. $SM(C, W)$ measures if C performs a number of emotion transitions similar to the actual number of emotion transitions in W. We use SM as a surrogate for short misclassifications. In the next section we introduce EC, an ensemble method that reduces the value of $T(C, W)$ by using a restrictive voting rule. Since in practice the value of $T(W)$ is small (the average is 7 in Padovani et al.'s dataset) and $T(C, W)$ tends to be larger than $T(W)$, by reducing $T(C, W)$, EC reduces $SM(C, W)$.

Ensemble Classifier (EC)

Given a set of classifiers C, EC classifies the emotion of the first sentence in the episode according to a majority voting rule of the classifiers in C; ties are broken randomly. For any other bag of words w, Bardo with EC only transitions from one emotion to another emotion e if all classifiers in C agree that w is of emotion e. That is, for the first sentence in the episode EC uses a majority voting rule and for every other sentence EC uses a unanimity voting rule. This unanimity rule is a special case of the rule introduced by Xu, Krzyzak, and Suen (1992). In Xu, Krzyzak, and Suen's rule a sample is "rejected" by the ensemble if all classifiers do not agree on the sample's label. In our case, if the classifiers do not agree on a label, we assume the emotion has not changed.

Theoretical Analysis

Our analysis is divided into two parts. First, we show that EC is expected to perform fewer emotion transitions than a single classifier. Then, we show sufficient conditions for EC to be more accurate than a single classifier in expectation.

Reduced Number of Transitions Let E be a set of emotions, C a set of classifiers used with EC, and $n = |C|$. We assume the probability of a classifier $c \in C$ classifying a bag of words w of emotion e_j as being of emotion e_i to be the same for all w of emotion e_j. We denote such probability as $p_c(e_i|e_j)$. Similarly, $p_C(e_i|e_j)$ is the probability of all classifiers in C classifying any w of emotion e_j as being of emotion e_i. We write p_c and p instead of $p_c(e_i|e_j)$ and $p_C(e_i|e_j)$ whenever e_i, e_j, and C are clear from the context.

Two events can occur in our problem: (i) EC correctly classifies the current emotion e_j or (ii) EC incorrectly classifies the current emotion e_j. In this part of the analysis we assume these events to be independent (i.e., the chances of a bag of words w_{t+1} being of emotion e is independent of the emotion of w_t). Assuming independence, the expected number of trials EC performs for event (i) to occur is given by $B(C, e_j) = p_C(e_j|e_j)^{-1}$. Similarly, the expected number of trials EC performs for event (ii) to occur is given by $R(C, e_j) = b_C(e_j)^{-1}$. Here, $b_C(e_j) = \sum_{\substack{e \in E \\ e \neq e_j}} p_C(e|e_j)$. $R(C, e_j)$ is the expected number of trials until all classifiers agree on an emotion different than e_j. We write b instead of $b_C(e_j)$ whenever e_j and C are clear from the context. Note that EC with $C = \{c\}$ is equivalent to c alone.

The following observation states that Bardo using EC with $n > 1$ is expected to change emotions less frequently than Bardo with any of its classifiers individually.

Observation 1 *For $C = \{c_1, c_2, \cdots, c_n\}$ and a subset of size one $C' = \{c\}$ with c being any classifier in C, we have that $B(C, e_j) \geq B(C', e_j)$ and $R(C, e_j) \geq R(C', e_j)$.*

If all classifiers in C are identical, $B(C, e_j) = B(C', e_j)$ and $R(C, e_j) = R(C', e_j)$, as the classifiers will always agree on the emotion transitions. If the classifiers in C are independent, then $p_C(e_i|e_j) = \prod_{c \in C} p_c(e_i|e_j)$ and the values of B and R will grow quickly with the size of C. Large B and R values mean that Bardo switches the background music less often (i.e., small $T(C, W)$ values). Since the number of emotion transitions $T(W)$ is small in practice, by

reducing $T(C, W)$ one is expected to reduce the value of $SM(C, W)$, our surrogate for short misclassifications.

Improved Overall Accuracy As one adds distinct classifiers into C, EC will require an increasingly larger number of trials before detecting an emotion transition. In particular, if the number of trials is larger than the number of sentences in a *scene*, then EC might miss the emotion transition entirely. A scene is an excerpt of a game session composed of bag of words with the same emotion.

Definition 1 (Scene) *Let $S = \{w_i, w_{i+1}, \cdots, w_{j-1}, w_j\}$ be a subset of W with $i \geq 1$ and $j \leq m$. Also, all $w \in S$ have the same emotion e and the emotions of w_{i-1} and w_{j+1} are different from e (if $i > 1$ and $j < m$); we call S a scene.*

The execution of EC within S can be modeled with two states: X and Y. EC is in X at time step t if it correctly identified the emotion of w_{t-1}. Since w_{t-1} and w_t have the same emotion (they belong to the same scene), if the classifiers do not agree on an emotion, then EC correctly classifies w_t by assuming it has the same emotion as w_{t-1}. EC is in Y at time step t if EC classified w_{t-1} as being of an emotion different from w_t's actual emotion. In our analysis we assume EC to start in Y. EC starts in X if the classification performed by EC correctly identifies the emotion of the first bag of words in the first scene of an episode, or if it misclassifies the last bag of words of a scene and the predicted emotion is the emotion of the next scene.

We define as q the size of a scene S and model the expected number of bag of words correctly classified by EC in S as $F_Y(q)$, which can be written with the recurrence:

$$F_Y(q) = p(F_X(q-1) + 1) + (1-p)F_Y(q-1) \quad (1)$$

$$F_X(q) = bF_Y(q-1) + (1-b)(F_X(q-1) + 1)\,. \quad (2)$$

Here, $F_Y(0) = 0$ and $F_X(0) = 0$, and p and b are the probabilities of EC correctly and incorrectly classifying the emotion of the current sentence of the scene, respectively. Function $F_Y(q)$ reads as "the number of bag of words EC is expected to correctly classify in the remaining q bags of the scene, given that EC is in state Y". Function $F_X(q)$ can be read similarly, except that it computes the expected number of bag of words classified correctly if EC is in state X.

Once a scene starts in Y, EC correctly classifies the current bag of words with probability p, thus adding one to the summation and transitioning to state X with $q-1$ bag of words remaining in the scene (see first term of $F_Y(q)$). EC misclassifies the current bag of word with probability $1-p$ and remains in state Y with $q-1$ bag of words remaining in the scene (see second term of $F_Y(q)$). Once in X, EC correctly classifies the remaining bags of words if the classifiers do not agree on an incorrect emotion (probability b). Equations 1 and 2 assume p and b to be the same for all w.

The following lemma shows that $F_Y(q)$ can be written as a closed-form equation. The proof is in the Appendix.

Lemma 1 *$F_Y(q)$ can be written as follows,*

$$\frac{p\Big((1-p-b)^{q+1} + p + b - 1 + q(p+b)\Big)}{(p+b)^2}\,.$$

$F_X(q)$ can be written as follows,

$$\frac{-b(1-p-b)^{q+1} + p^2q + pbq - pb - b^2 + b}{(p+b)^2}\,.$$

Lemma 1 allows us to derive the minimum size q of a scene to guarantee that EC is expected to be more accurate than a single classifier with accuracy k.

Theorem 1 *Let S be a scene of size $q \geq 0$ and c a classifier with accuracy $k \in (0, 1]$ in S. Assuming that the probability values $p, b \in (0, 1]$ are fixed for all bag of words in S, EC is more accurate than c if $q > \frac{p^2-p+pb}{(p+b)^2k-p^2-pb}$ and $k < \frac{p}{p+b}$.*

The proof of Theorem 1 is in the Appendix. Theorem 1 states that if S is long enough and c is not too accurate, then EC is expected to be more accurate than c in S. Note that a regular classifier is a special case of EC with an ensemble of size one. In that case, $b = 1 - p$, which according to Theorem 1, $q > 0$ as long as $k < p$, as one expects.

Our theoretical results suggest that EC is able to reduce short misclassifications and can be more accurate than a single classifier. On the other hand, EC might miss the emotion transitions of short scenes. This is because short scenes might finish before EC transitions from state Y to state X.

Although EC uses the restrictive unanimity rule, our analysis holds for other voting rules such as the majority rule. In that case, p and b mean the probability of the majority of the classifiers in the ensemble classifying a bag of words correctly or incorrectly, respectively. We chose to use the unanimity rule because this rule is expected to result in larger values of B and R, which can potentially reduce the sparsely distributed errors and thus allow us to test empirically our hypothesis that it might be preferred to be somewhat inaccurate and decisive than accurate but often indecisive. Also, note that one could also analyze EC by treating it as a Markov Chain with states X and Y whose transition matrix is defined by p, $p-1$, b, and $b-1$.

Empirical Evaluation

In this section we evaluate variants of EC and Naive Bayes (NB) on the 9 episodes of CotW. The goal of this experiment is to show empirically that EC is able to reduce the number of short misclassifications and is thus suitable to test our hypothesis that short misclassifications can harm how the user perceives their experience with the system.

NB classifies a bag of words w according to the probability of each word in w belonging to a class and according to the a priori probability of a sentence belonging to a class (Manning et al. 2008). Two of the NB models we use are created by choosing different sliding window sizes z. We use a leave-one-episode-out cross-validation procedure in the set of training episodes to select the two sizes. In the leave-one-episode-out cross-validation procedure we remove one episode from the set of training episodes and train the model on the remaining episodes. The model is then evaluated on the held-out training episode. This process is repeated for all possible episodes in the training set. One NB model is obtained by selecting the sliding window size that yields the model with largest average accuracy in

Alg.	Episodes									Avg.
	1	2	3	4	5	6	7	8	9	
Accuracy										
Baseline	10	53	35	44	75	64	29	24	47	42
NS	64	62	76	71	54	69	44	59	79	64
NHS	71	57	79	69	56	59	55	59	76	64
NM	64	62	80	72	52	60	43	61	78	64
NHM	68	57	79	69	47	65	56	61	76	64
EC(2)	71	60	78	69	54	59	56	59	81	65
EC(4)	76	59	79	70	50	59	55	64	80	65
SM										
Baseline	7	8	5	12	8	9	4	6	4	7
NS	42	29	43	40	37	41	59	35	36	40
NHS	45	28	23	33	35	22	50	42	50	36
NM	40	24	39	27	45	30	58	31	36	36
NHM	47	28	19	29	33	37	49	25	50	35
EC(2)	14	4	4	10	9	0	25	10	13	9
EC(4)	5	3	4	6	5	0	12	2	11	5

Table 1: Accuracy and SM for different classification algorithms.

the cross-validation procedure; we call this model NS. The other model, called NM, is defined similarly, but by selecting the sliding window size that yields the model with lowest average SM. We test windows with size: $\{20, 25, 30, 35, 40\}$. The classifier NS is identical to the one used by Padovani, Ferreira, and Lelis (2017).

We create two extra NB models by using a feature selection scheme. A NB model with feature selection uses only the h words with largest mutual information (MI) value (Manning et al. 2008) in its classification procedure. The MI value of a word measures how discriminative the word is to identify or to rule out a given class. We then create NHS and NHM, which are similar to NS and NM, except that in the cross-validation procedure we choose the value of z and h that maximizes accuracy (for NHS) and minimizes SM (for NHM). We test the following values of h: $\{500, 600, \cdots, 1500\}$, resulting in a total of 55 values tested for NHS and NHM. We test two versions of EC, one with NS and NHS composing its ensemble (EC(2)) and one with NS, NHS, NM, and NHM in its ensemble (EC(4)).

Since the number of transitions is small in CotW and EC tries to minimize SM by reducing the number of emotion transitions, a reasonable baseline is to assume Bardo plays as background music a song related to the Suspenseful emotion, which is the majority class in Padovani et al.'s dataset. We call this approach Baseline in our table of results.

We separate each episode to be tested and train the algorithms on the other episodes. For example, when testing a method on episode 1, we train it with episodes 2–9 and the resulting model is applied to episode 1.

Accuracy and SM Results

Table 1 shows the percentage accuracy (upper part) and SM values (bottom part) of the tested algorithms ("Alg.") in each episode of CotW. The "Avg." column shows the algorithm's average results across all episodes. All numbers are rounded to the closest integer and the values in the "Avg." column were computed before rounding the numbers. We highlight the background of a cell if the number in the cell represents the best result across all algorithms for a given episode. For example, NM has the has the highest accuracy in episode 3 and EC(4) has the lowest SM value in episode 4. We also highlight the best overall averages for each episode.

The EC approaches present a much lower SM than all classifiers. Namely, EC(4)'s SM is 7 times lower than NHM's, the best performing individual classifier. EC(4) also has an average SM lower than Baseline and a much higher accuracy. As suggested by our analysis, EC with a larger set of classifiers is expected to change the emotion less often, potentially further reducing the value of SM. EC(2) and EC(4) have the same average classification accuracy, but EC(4)'s SM is nearly half of the SM value of EC(2).

The SM reduction performed by EC(4) can be observed in Figure 1, which shows the EC(4) classifications in green in an excerpt of episode 6. EC(4) has a SM of 3 in the excerpt shown in Figure 1: there are 8 true emotion transitions shown in red while EC performs 5 transitions. The SM value of EC(4) is zero if one considers the entire episode (Episode 6 in Table 1). As anticipated in our theoretical analysis, EC's misclassifications are not sparsely distributed, they usually happen at the beginning of a scene. For example, there is a change from Calm to Suspenseful around sentence 500 (see red lines in Figure 1) and EC only detects such a transition after approximately 20 sentences. Also, EC misses entirely some of the short scenes in the episode (e.g., the transitions in between sentences 400 and 500). The presence of short scenes justify the individual classifiers being more accurate than EC in this episode (see Table 1), as suggested by Theorem 1.

User Study

The results presented so far show that EC is able to change the error distribution of its base classifiers. Instead of erring sparsely throughout an episode, EC tends to concentrate its errors in the beginning of the scenes, thus reducing the short misclassifications. In this section we test with a user study our hypothesis that the sparsely distributed errors can harm how the user perceives their experience.

Our system for background music selection can target two types of users: people playing the game (players) and people watching the game session (spectators). A player would need to commit a few hours to participate in our study. By contrast, spectators need to commit only a few minutes to watch video excerpts of game sessions. Since it is easier to enlist spectators than players, our study focus on measuring approximations of the spectator's perceived experience. We approximate the spectator's perceived experience with a set of pairwise preference comparisons of the background music selected by EC and by NS, Bardo's original classification model, in video excerpts of people playing D&D.

Our study is designed such that the results obtained with evaluations of short video excerpts are likely to generalize to longer game sessions. This is achieved by selecting video excerpts with different accuracy and SM values so

that the excerpts cover a variety of scenarios that might arise in longer game sessions. Also, we do not use the baseline used in the previous experiment (play a song of the majority class throughout the excerpt). This is because while such a baseline might yield reasonable results in short excerpts, the results are unlikely to generalize to longer game sessions as playing the same type of song will likely bore the users and harm their perceived experience. Moreover, Padovani, Ferreira, and Lelis (2017) showed that NS is able to outperform in terms of user preference a much stronger baseline, which is the background music selected by the authors of the videos used in our user study.

We compare NS with EC(4) (henceforth referred as EC). We selected five excerpts of the CotW. NS and EC are trained with episodes different than the one from which the excerpt is extracted (e.g., if an excerpt is extracted from episode 7, then NS and EC are trained with all episodes but 7). Each excerpt is approximately 2 minutes long. In order to test our hypothesis, we selected excerpts in which NS is either as accurate or more accurate than EC and has a larger SM value, thus giving NS an advantage in terms of system accuracy. The accuracy and SM values of the video excerpts (V1, V2, V3, V4, V5) are shown at the bottom of Table 2.

The video excerpts we use have no sentences of the Happy emotion (Happy is a rare emotion in CotW), thus we use one song for each of the other emotions in our study. We used the song *Call of the Raven* by Jeremy Soule for Calm, *Hurricane Suite* by Naruto Shippuden OST I for Suspenseful and *Open the Gates of Battle* by Casey Martin for Agitated. V1 is an excerpt starting at 3:22 and finishing at 5:20 of episode 2 of CotW; V2 starts at 25:10 and finishes at 26:32 of episode 6; V3 starts at 23:26 and finishes at 24:55 of episode 4; V4 starts at 17:26 and finishes at 18:56 of episode 3; V5 starts at 20:00 and finishes at 21:31 of episode 7.

Following Padovani et al.'s methodology, each participant listened to excerpts of all three songs after answering our consent form and before evaluating the video excerpts. We reduce the chances of a participant evaluating the quality of the songs instead of the song selection procedure by telling the participant which songs will be used as background music. After listening to the songs each participant watched two versions of the same video excerpt, one with the background music selected by NS and another by EC. The order in which the videos appeared was random to avoid ordering biases. We included a brief sentence providing context to the participant, to ensure they would understand to story being told in each excerpt. The participants could watch each video as many times as they wanted before answering the question: "Which video has the most appropriate background music according to the context of the story?". The participant could choose one of the options: "Video 1", "Video 2", "The background music used in both videos are equally appropriate", and "The background music used in both videos are equally inappropriate". After marking their answer, the participants evaluated another pair of excerpts. The order the video pairs were presented was also random. The participants answered a demographic questionnaire after evaluating all excerpts.

Our experiment was advertised in D&D communities in the social media. We had 40 participants, 39 males and 1 female, with average age of 25. All participants had some experience playing D&D. We report the results of 200 answers (5 pairs of videos for each participant).

Method	Video Excerpts				
	V1	**V2**	**V3**	**V4**	**V5**
EC	57.5	57.5	35.0	35.0	47.5
NS	10.0	12.5	37.5	37.5	22.5
Tie+	17.5	12.5	15.0	22.5	20.0
Tie-	15.0	17.5	12.5	5.0	10.0
Accuracy					
EC	87.8	0.0	32.3	48.3	32.3
NS	85.7	20.0	41.9	69.0	38.7
SM					
EC	0	0	0	1	1
NS	6	2	10	3	9

Table 2: User preference in emotion detected by EC and NS.

User Study Results

The videos with background music selected by EC were preferred 93 times by the participants, while NS's videos were preferred 48 times, and the approaches tied 59 times. The difference between EC and NS is significant according to a two-sided binomial test ($p < 0.001$).

Table 2 shows the detailed results for all 5 excerpts used in the study. The upper part of the table shows the percentage of times the participants preferred the videos edited by EC, by NS, and the percentage of times the participants thought the videos to be equally appropriate (Tie+), and equally inappropriate (Tie-). For example, for the first two excerpts (V1 and V2), the participants preferred EC's selection of background music in 57.5% of the cases. The highlighted cells show the best performing approach (EC or NS) on a given excerpt. The bottom part of the table shows EC and NS's accuracy and the SM value in each excerpt.

The results of our user study show a clear preference for the music selected by EC. In particular, the participants strongly preferred the selection performed by EC in V1, V2, and V5. V1 is an excerpt with two scenes in which both methods are accurate. While NS's misclassifications are sparsely distributed in V1, EC's misclassifications occur in the beginning of one of the scenes due EC requiring a few sentences to detect the emotion change.

EC classified all sentences in V2 as Agitated while the sentences were of the Calm emotion. NS correctly selected the Calm song for part of the excerpt but switched a few times between Calm and Agitated. In this case, the participants preferred the selection that was inaccurate and decisive over the selection that was more accurate but indecisive. EC performs similar misclassification in V3, where it selects the Agitated song for a Suspenseful scene. In contrast with V2, EC's misclassifications in V3 were not well perceived by the participants. V3 depicts a scene in which one of the players

is sneaking in their enemy's house. The use of an Agitated song instead of a Suspenseful seems to be more harmful to the user's perceived experience than a large SM value in this particular case. Note, however, that that the participants only marginally prefer the selections made by NS in V3. A similar result is observed in V4, where EC is less accurate than NS but has a lower SM value and participants only marginally prefer NS's selections. The accuracy of EC and NS are similar in V5, but the latter has a large SM value. In this case the participants have a strong preference for EC's selections.

The study supports our hypothesis that sparsely distributed errors can be harmful to the user's perceived experience. The study also showed that in some cases it might be preferable to be inaccurate and decisive than accurate and indecisive (e.g., V2). Naturally, depending on the scene, the lack of accuracy can outweigh the system's decisiveness (e.g., V3). Video excerpts V1 and V5 are the most representative videos used in the study. This is because NS and EC are equally accurate (accuracy of approximately 60% considering V1 and V5 altogether) and EC has a much smaller SM value; Table 1 showed similar average accuracy results for EC. The participants showed a strong preference for EC's selection in V1 and V5. Thus, overall, our results suggest that EC is the ideal method to be used for selecting background music as it is able to reduce the SM value without sacrificing the accuracy of its base models.

Conclusions and Future Work

In this paper we studied how the error distribution of a system can affect the user perceived experience. We hypothesized that a system whose errors are sparsely distributed across a game session could harm the user perceived experience and that it would be preferred to be somewhat inaccurate but decisive than accurate and indecisive. In order to test our hypothesis, we introduced an ensemble approach called EC that errs consistently in the beginning of scenes as opposed to erring sparsely through the scenes. Theoretical results showed that EC can reduce the sparsely distributed errors by performing fewer transitions. We also showed that if a scene S is long enough and a classifier c is not too accurate, then EC is expected to be more accurate than c in S. Empirical results showed that EC is able to reduce short misclassifications without sacrificing accuracy. A user study in which people watched videos of D&D with the background music selected by EC and by a classifier that does not account for the error distribution supported our hypothesis.

Although we evaluated our system with an ensemble of classifiers in the domain of background music selection, our results might generalize to other classifiers and application domains. We expect future works to study how other classifiers can be modified to cope with short misclassifications and how a system's error distribution can affect user experience in other application domains. Also as future work, we are interested in verifying if the findings of our study indeed generalize to longer sessions. Longer experiments will also allow us to measure the impact of error distribution with concrete metrics of user experience such as churn rate (instead of the user preference metric used in our study).

Appendix: Proofs

Lemma 1 *$F_Y(q)$ can be written as follows,*

$$\frac{p\Big((1-p-b)^{q+1}+p+b-1+q(p+b)\Big)}{(p+b)^2}.$$

$F_X(q)$ can be written as follows,

$$\frac{-b(1-p-b)^{q+1}+p^2q+pbq-pb-b^2+b}{(p+b)^2}.$$

Proof. Our proof is by induction. Replacing $q=0$ in the equations above we obtain $F_Y(0)=p(1-p-b+p+b-1)=0$ and $F_X(0)=-b+bp+b^2-pb-b^2+b=0$.

We assume as inductive hypothesis (IH) that $F_Y(q-1)=p\big((1-p-b)^q+p+b-1+(q-1)(p+b)\big)(p+b)^{-2}$ and $F_X(q-1)=\big(-b(1-p-b)^q+p^2(q-1)+pb(q-1)-pb-b^2+b\big)(p+b)^{-2}$.

By replacing $F_Y(q-1)$ and $F_X(q-1)$ according to the IH in the recursive version of $F_y(q)$ we obtain its closed form, as stated in the lemma. Similarly, one obtains the closed-form version of $F_X(q)$ by replacing $F_Y(q-1)$ and $F_X(q-1)$ according to the IH in the recursive version of $F_X(q)$. □

Theorem 1 *Let S be a scene of size $q\geq 0$ and c a classifier with accuracy $k\in(0,1]$ in S. Assuming that the probability values $p,b\in(0,1]$ are fixed for all bag of words in S, EC is more accurate than c if $q>\frac{p^2-p+pb}{(p+b)^2k-p^2-pb}$ and $k<\frac{p}{(p+b)}$.*

Proof. EC is expected to be more accurate than C if

$$\frac{p\big((1-p-b)^{q+1}+p+b-1+q(p+b)\big)}{(p+b)^2}>kq$$

$$\frac{(p+b)^2kq}{p}-p-b+1-q(p+b)<(1-p-b)^{q+1}$$

Since $(1-p-b)^{q+1}\geq 0$, the equation above holds if

$$\frac{(p+b)^2kq}{p}-p-b+1-q(p+b)<0 \tag{3}$$

$$q\big((p+b)^2k-p^2-pb\big)<p^2-p+pb \tag{4}$$

$$q>\frac{p^2-p+pb}{(p+b)^2k-p^2-pb} \tag{5}$$

In Equation 5, p^2-p+pb is negative as one needs $b+p>1$ for it to be positive, and $b+p\leq 1$. Thus, Equation 5 holds if $(p+b)^2k-p^2-pb<0$, or $k<\frac{p}{(p+b)}$. Suppose $(p+b)^2k-p^2-pb>0$, since $p^2-p+pb<0$, then one needs $q<0$ for Equation 4 to hold, but $q\geq 0$. □

Acknowledgements This research was supported by CNPq (200367/2015-3) and FAPEMIG. We thank the participants of our user study, in special those from the group D&D Next. We thank the Node YouTube channel for allowing us to use their videos in our research. We also thank Sandra Zilles for suggesting a proof by induction for Lemma 1 and the anonymous reviewers for other great suggestions.

References

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Cherkauer, K. 1996. Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, 15–21.

Chou, S.-C. T.; Yang, C.-C.; Chan, C.-H.; and Lai, F. 1996. A rule-based neural stock trading decision support system. In *IEEE Conference on Computational Intelligence for Financial Engineering*, 148–154.

de Vries, P.; Midden, C.; and Bouwhuis, D. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *Int. J. Hum.-Comput. Stud.* 58(6):719–735.

Desai, M.; Medvedev, M.; Vázquez, M.; McSheehy, S.; Gadea-Omelchenko, S.; Bruggeman, C.; Steinfeld, A.; and Yanco, H. 2012. Effects of changing reliability on trust of robot systems. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 73–80. New York, NY, USA: ACM.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–15. London, UK, UK: Springer-Verlag.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119–139.

Hansen, L. K., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions Pattern Analysis and Machine Intelligence* 12(10):993–1001.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Lee, J. D., and See, K. A. 2004. Trust in automation: designing for appropriate reliance. *Human Factors* 46(1):50–80.

Manning, C. D.; Raghavan, P.; Schütze, H.; et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Padovani, R.; Ferreira, L. N.; and Lelis, L. H. S. 2017. Bardo: Emotion-based music recommendation for tabletop role-playing games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Sanchez, J. 2006. *Factors that affect trust and reliance on an automated aid.* Ph.D. Dissertation, Georgia Institute of Technology.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning* 5(2):197–227.

Seo, Y.-W.; Giampapa, J.; and Sycara, K. 2004. Financial news analysis for intelligent portfolio management. Technical report, Robotics Institute, Carnegie Mellon University.

Xu, L.; Krzyzak, A.; and Suen, C. Y. 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics* 22(3):418–435.

Yang, X. J.; Unhelkar, V. V.; Li, K.; and Shah, J. A. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 408–416. New York, NY, USA: ACM.