A NOVEL TOOL FOR IMPROVING THE DATA COLLECTION PROCESS DURING CONTROL ROOM MODERNIZATION HUMAN-SYSTEM INTERFACE TESTING AND EVALUATION ACTIVITIES

Kovesdi, C., Joe, J.
Idaho National Laboratory
PO Box 1625, Mail Stop, 3818 Idaho Falls, Idaho 83415, USA
Casey.Kovesdi@inl.gov; Jeffrey.Joe@inl.gov

ABSTRACT

The U.S. Department of Energy (DOE) Light Water Reactor Sustainability (LWRS) Program is developing a scientific basis to extend the existing U.S. nuclear power plant (NPP) operating life beyond the current 60-year licensing period and to ensure NPP long-term reliability, productivity, safety, and security. Under the Advanced Instrumentation, Information, and Control (II&C) Systems Technologies pathway, the LWRS Program conducts targeted research and development to address long-term aging and obsolescence of existing instrumentation and control technologies, as well as to implement and test new digital II&C technology that enables broad innovation and business improvement in the NPP operating model. To this end, the role of human factors engineering (HFE) is important as part of ensuring that the control room modernization (CRM) activities account for the operator's cognitive and physiological characteristics. Part of the HFE process for CRM entails conducting tests and evaluations (T&Es) that will inform the design of new human-system interface (HSI) technologies. Qualitative data that captures operator feedback and important human-system interactions are an important part of T&E. However, the analyzing this qualitative data in order to provide actionable design recommendations can be time-consuming and error prone. This paper presents an early version of a tool, which uses a combination of natural language processing and supervised machine learning to support the analysis of qualitative data. The objectives of this tool is to improve the process of (1) sifting through freeform text for improved design issue identification, (2) maintaining consistency in theme coding conventions, and (3) readily providing actionable design guidance that can be traceable to state-of-the-art HFE guidelines such as NUREG-0700.

Key Words: control room modernization, human factors engineering, tests and evaluations, qualitative data analysis

1 INTRODUCTION

The United States (U.S.) total energy consumption is expected to increase from 2016 to 2040, with the commercial sector being a significant source of electrical consumption [1]. To support the U.S. electricity demands with a reliable and economically viable resource, the U.S. Department of Energy (DOE) Light Water Reactor Sustainability (LWRS) Program is developing a scientific basis to extend the existing U.S. nuclear power plant (NPP) operating life beyond the current 60-year licensing period and to ensure NPP long-term reliability, productivity, safety, and security. The LWRS Program accomplishes this objective through a multi-pathway approach of targeted research and development (R&D) focus areas. One R&D focus area, the Advanced Instrumentation, Information, and Control (II&C) Systems Technologies pathway, conducts targeted R&D to address long-term aging and obsolescence of existing instrumentation and control (I&C) technologies, as well as to implement and test new digital II&C technology that enables broad innovation and business improvement in the NPP operating model.

Human factors engineering (HFE) plays a major role in the Advanced II&C Systems Technologies pathway through ensuring that the control room modernization (CRM) activities account for the cognitive and physiological characteristics of the operator. The U.S. Nuclear Regulatory Commission (NRC) recognizes the importance of HFE in supporting plant safety and providing defense in depth [2]. An applicant (i.e., utility) is expected to have a HFE program put in place to which the NRC reviews in order to verify that the applicant's HFE program incorporates state-of-the-art HFE practices and guidelines accepted by the NRC staff. This HFE review model is described in the *Human Factors Engineering Program Review Model* (NUREG-0711, Rev. 3).

One of these elements in NUREG-0711, Human-System Interface (HSI) Design, is important to CRM such that human engineering deficiencies (HEDs) and HSI design issues are identified and corrected earlier in the development process to ensure later-stage elements are successful. Indeed, one motive for this early focus of HFE is to substantially reduce costs and save time during HSI development [3]. NUREG-0711 describes these early HSI design activities used to identify and correct HEDs and design issues Tests and Evaluations (T&Es). During T&E, different HSI design concepts (i.e., prototypes) can be explored and evaluated using actual users of the HSI (i.e., plant personnel).

T&Es during HSI Design are less rigid and are informal compared to Verification and Validation (V&V) activities. Recently, Boring and colleagues [4] developed a process for CRM that utilities can follow to maintain consistency across multiple system upgrades termed as the *Guideline for Operational Nuclear Usability and Knowledge Elicitation* (GONUKE). GONUKE describes three types of evaluation that can be completed across each NUREG-0711 phase including expert review, user study, and knowledge elicitation. Expert review is defined as the evaluation of a system, by a subject matter expert, against a standardized set of evaluation criteria. User study is the evaluation of human-system performance. Finally, knowledge elicitation is the process of capturing insights of the users who use the system as pertained to the HSI design. Moreover, the need for *qualitative input* from plant personnel collected during T&E is emphasized in this process as a tool to guide HSI Design, especially with knowledge elicitation.

This qualitative data is generally captured as freeform textual data. The process of capturing qualitative data demands full attention of the data recorder and favors minimizing the number of extraneous tasks (e.g., tasks other than recording notes) during data collection. The consequence of missing information during qualitative data collection can result in incorrect conclusions or missing potential design issues. Common qualitative methods used during CRM-T&E include collecting freeform notes via pen and paper or electronically using a spreadsheet. The advantage of using basic tools such as these pertains to the flexibility and ease of data recording. For example, these basic methods are not as influenced by last minutes changes to a study protocol as a programmed web-based form might be.

However, an important potential tradeoff with having more flexibility with freeform notes is the lack of consistency placed on coding HSI design issues and parsing ancillary information from true design issues. This pitfall in the current data collection process ultimately requires additional effort in distilling key themes observed during T&E from the notes taken. Inter-rater reliability, particularly with missing potential design issues from notes, is a concern [5]. Additionally, there is an additional layer of effort translating comments and observations made during a CMR-T&E activity into actionable design recommendations that can be traceable to state-of-the-art HFE guidelines such as from the U.S. NRC's *Human-System Interface Design Review Guidelines* (NUREG-0700, Rev. 2). For example, NUREG-0700 Rev. 2 contains nearly 2200 different guidelines that the HFE practitioner must consider when providing design guidance.

This paper presents an early version of a tool that can be used to analyze the qualitative data collected during HSI Design CRM-T&Es to (1) sift through freeform text for improved design issue identification, (2) maintain consistency in theme coding conventions, and (3) readily provide actionable

design guidance that can be traceable to state-of-the-art HFE guidelines such as NUREG-0700. A description of this tool is discussed next.

2 DESCRIPTION OF THE QUALITATIVE DATA ANALYSIS TOOL

This qualitative data analysis tool was built using the open source R framework [6] and uses several of the packages offered from the R-CRAN repository such as: cluster [7], fpc [8], hunspell [9], koRpus [10], lsa [11], NLP [12], openNLP [13], qdap [14], reshape [15], SnowballC [16], stringr [17], tcltk2 [18], tm [19], and XLConnect [20]. These packages enable graphical user interface (GUI) development, as well as leverages R's capabilities in text processing and analytics, natural language processing (NLP), and machine learning (ML). Fig 1 illustrates the tasks and functions that the tool and user perform to provide actionable design recommendations.

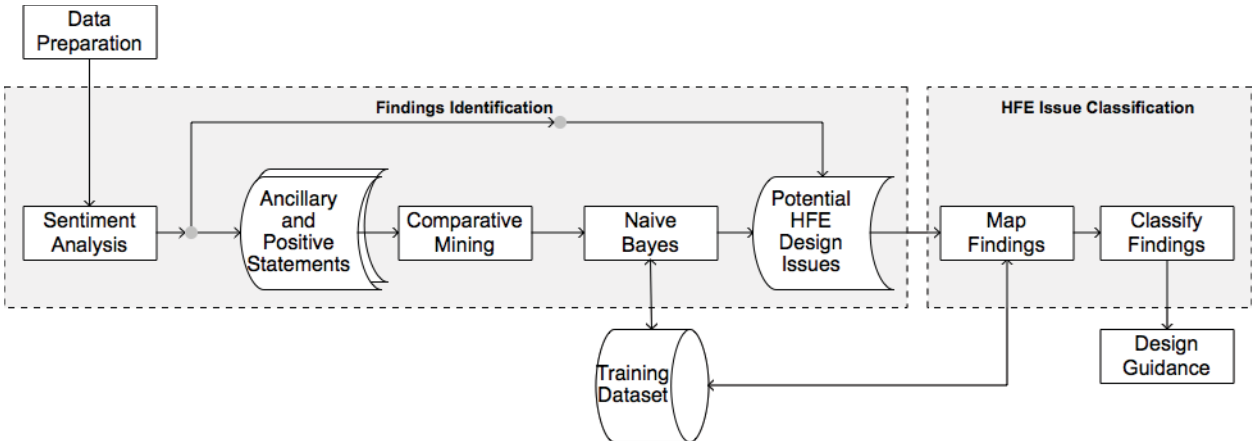

Figure 1. Workflow of the HFE Classification Tool.

2.1 Data preparation

Qualitative data collected during CRM-T&E first must be imported into the tool for subsequent processing. The tool reads freeform notes stored in a particular column of a Microsoft Excel (i.e., .xls or .xlsx) spreadsheet using the XLConnect [20] package. The user is then able to (1) select the Excel file of interest, (2) select the worksheet within the Excel file, and finally (3) select the column of the freeform notes. The tool then completes a spellcheck using the Hunspell algorithm from the hunspell [9] package. This step is necessary in order to accurately complete subsequent processes involving sentiment analysis and NLP.

Once the spellcheck is complete, freeform text must be grouped meaningfully such that a complete statement (i.e., a sentence) is presented. This step is pertinent for subsequent text analysis as a single note may include multiple statements. The tool identifies sentences by mapping each word's part of speech (POS) from the NLP [12] and openNLP [13] packages. These R packages interface with the Apache OpenNLP library [21], an open source ML based toolkit, for common NLP tasks such as POS tagging. Identified sentences from POS tagging are stored as a new dataset in R and used in additional text analyses described next in *Findings Identification*.

2.2 Findings Identification

This section describes the methods and analyses used to classify statements into potential design issues. The philosophy for identifying potential design issues places greater costs of misclassifying a design issue as a non-issue (i.e., miss) compared to misclassifying a non-issue as a design issue (i.e., false

alarm). The tool provides all potential design issues to the user to ultimately decide whether the statement refers to a design issue or not. Prematurely ruling out statements as potential design issues will result in failure of providing the user an opportunity to make such a decision. This functionality by the tool should reduce the number of statements that are irrelevant to HFE interface design, which may have been recorded for other purposes.

2.2.1 Sentiment Analysis

The first analysis used to identify potential design issues entails sentiment analysis, which uses the qdap package [14]. The sentiment analysis algorithm uses a sentiment dictionary from Hu and Liu [22], and assigns a sentiment value (δ) to each sentence by tagging each individual word as a neutral (t^0), negator (t^N), amplifier (t^a), or de-amplifier (t^d) term. A neutral tag denotes terms without any positive or negative sentiment. While neutral words do not provide a polarity value, these words affect the total word count for a sentence (n). Negators are terms that influence sentiment as being positive or negative polarity. Amplifiers are terms that increase the intensity of being positive or negative sentiment, whereas de-amplifiers are terms that decrease the intensity of being positive or negative sentiment. A negator weight (ω_{neg}) is also created from the remainder the total negators divided of two. The following set of equations express how the algorithm assigns sentiment to each sentence using each of these context tags [i.e., see 14].

$$\delta = \frac{\sum\left[\left(1+0.8\left(t^A - t^D\right)\right)\cdot(-1)\sum t^N\right]}{\sqrt{n}} \tag{1}$$

$$t^A = \sum\left(\omega_{neg} \cdot t^a\right) \tag{2}$$

$$t^D = \max\left(\sum\left(-\omega_{neg} \cdot t^a + t^d\right), -1\right) \tag{3}$$

$$\omega_{neg} = \left(\sum t^N\right) \bmod 2 \tag{4}$$

The unbounded sentence sentiment values (δ) are then trichotomized as being negative (-1), neutral (0), or positive (+1), depending on whether the sentiment value is less than 0, equal to 0, or greater than 0. Sentences in each of these three sentiment classes are grouped into individual datasets: potential design issues (-1), ancillary statements (0), and positive statements (+1).

Identified potential design issues are provided to the user during HFE Issue Classification. However, this tool uses a conservative approach, which assumes that sentiment analysis may not uncover all statements that could contain potential design issues. For example, a neutral statement with a potential design issue may be, "*The font was not legible due to its contrast.*" This statement would be missed by sentiment analysis since no specific word from the statement yields positive or negative polarity. Likewise, a seemingly positive statement may implicitly reflect a potential design issue; see the statement, "*The indications in the existing control room are easier to see than the new interface*." Thus, ancillary and positive statements are considered for subsequent text analyses including comparative mining and Naïve Bayes classification. The next two sections describe methods used to additionally classify statements such as the ones illustrated in the two examples as potential design issues.

2.2.2 Comparative Mining

Comparative or superlative statements from the ancillary and positive statement datasets are automatically assumed to be potential design issues. To reiterate, the identification of a potential design issue does not ultimately determine whether the statement pertains to a design issue. Rather, these statements are presented to the user in order for him/ her to ultimately decide whether a design issue is

reflected within the statement. The tool classifies comparative or superlative statements by determining whether a comparative adjective (JJR), superlative adjective (JJS), comparative adverb (RBR), or superlative adverb (RBS) was present within the statement [23]. For example, the term 'easier' would be identified as a JJR from, "*The indications in the existing control room are easier to see than the new interface*." As a result, this statement would be classified as a potential design issue as shown in Fig. 2 below.

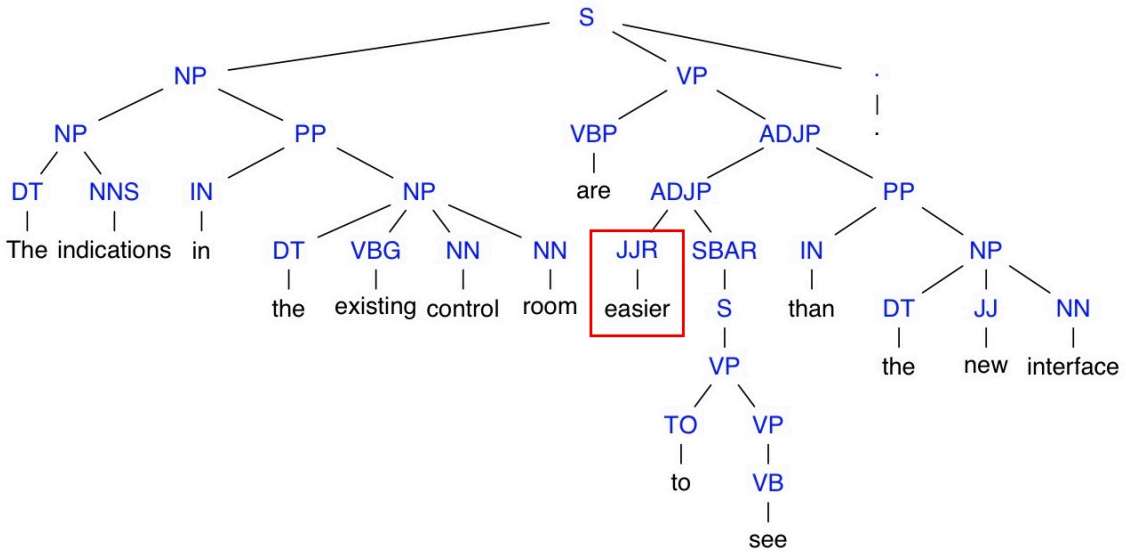

Figure 2. Constituent Structure of a Comparative Statement.

2.2.3 Naïve Bayes

Lastly, all ancillary and comparative statements are classified as being either a potential design issue or a non-issue using the supervised ML approach Naïve Bayes classification. The use of Naïve Bayes for text classification is an algorithm widely used for text classification problems such as with anti-spam email filtering [e.g., 24]. One motivation for using Naïve Bayes is that it is effective with noisy data (i.e., even when its assumptions are broken) and requires relatively few examples for training [25]. The Naïve Bayes classification algorithm computes a posterior probability that a statement is likely to be either a potential design issue or not using a training dataset of similar statements. For each statement, posterior probabilities (P_L) of each class level (C_L) are computed given the evidence provided from each key term ($t_{1,\dots,}\ t_n$), as expressed below.

$$P_L(C_L|\ t_1, \dots, t_n) = p(C_L) \prod_{i=1}^{n} p(t_i|C_L) \tag{5}$$

As illustrated in this equation, posterior probabilities are effectively a function of the product of the probabilities (p) of each evidence term conditional on the class level ($t_i\ |C_L$), and the prior probability of the class level (C_L) [26]. This tool also applies a Laplace correction (i.e., initializing each word count to 1 rather than 0) to account for the zero-frequency problem (i.e., accounting for terms in a statement that do not occur in the training dataset for a given class) as described in [26].

The initial Naïve Bayes classifier training dataset was developed from three previous HFE T&E workshops. Issues were manually coded by the HFE expert who originally recorded the data. Moreover, this training dataset is dynamic in nature such that the user is able to refine and build upon the existing dataset when deciding whether a statement is a design issue or not. The user is able to decide whether the

statement (e.g., "*The font was not legible due to its contrast.*") is a design issue or a non-issue from the tool's GUI. The classification of each statement by the user effectively is added to the training dataset to further improve the Naïve Bayes classifier's accuracy. A final point worth mentioning is that typically classification from posterior probabilities can be less concerned with biases in probability estimates (e.g., whether by 55% versus 95% confident in a prediction). However given that the philosophy of this tool is to be less apt to generate a miss compared to a false alarm, a different strategy is implemented. That is, the estimated confidence in predicting a non-issue needs to be at least four times more likely (i.e., ! 80% confident of being a non-issue) than a potential design issue; otherwise, the tool automatically assumes that the statement contains a potential design issue. Fig. 3 illustrates the tool's GUI where the user is able to decide whether the statement is a design issue (i.e., by selecting '7 of 7 >') or a non-issue (i.e., by selecting 'No Issue').

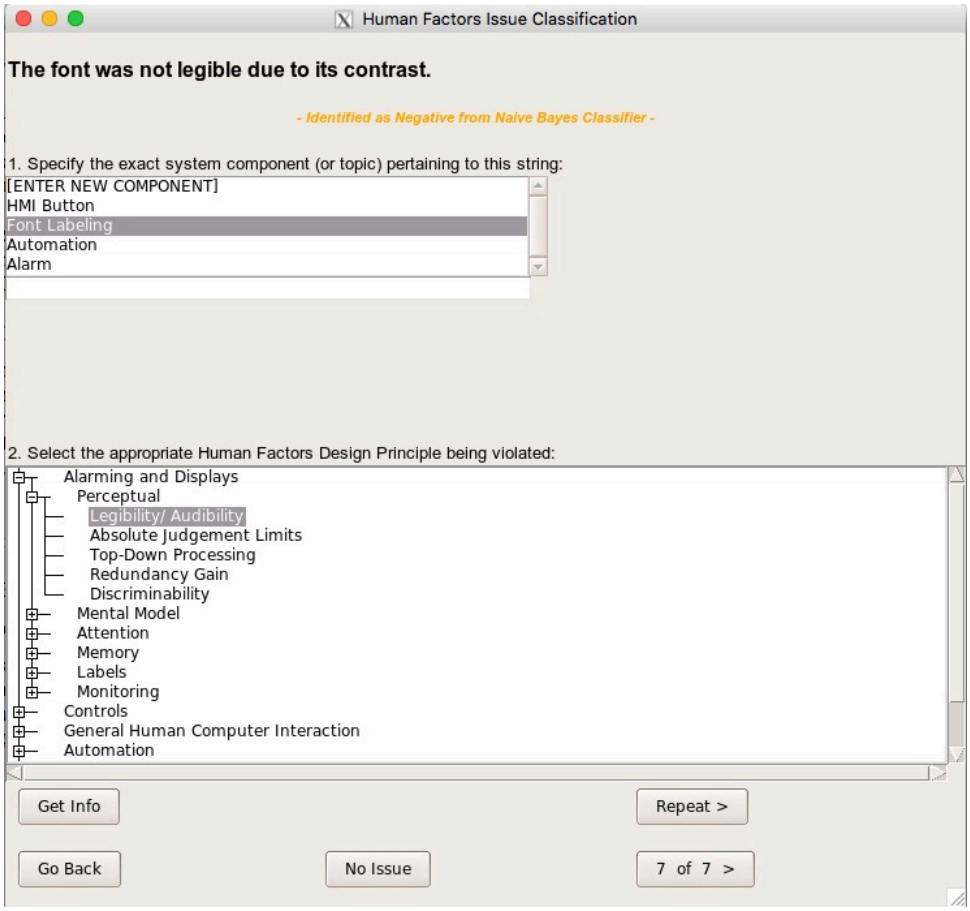

Figure 3. HFE Classification Tool Graphical User Interface.

2.3 HFE Issue Classification

Once the tool has completed data preparation and findings identification, the final task is to present to the user each statement with a potential design issue so that the user can (1) specify the exact system component or topic from each statement and (2) map each statement to a generalizable HFE design principle. As shown in Fig. 3, each statement is displayed at the top of the GUI. The GUI also displays the source of how the statement was identified (e.g., Sentiment Analysis, Comparative Mining, or Naïve Bayes classification). This task is completed for each and every statement with a potential design issue sequentially. The submit button (i.e., '7 of 7>') displays how many statements were identified in relation

to what statement the user is currently classifying (e.g., statement 7 of 7). Each generalizable HFE design principle is linked to NUREG-0700 guidelines that can be applicable to the design issue at hand. These design recommendations are provided to the user as a separate Microsoft Excel file where the user can pick and choose what guidelines are relevant.

2.3.1 Map Findings

For this step, the user is tasked to enter a system component (e.g., HMI button) or topic (e.g., Font Labeling) that applies to the statement at hand. The GUI allows the user to enter a new component/ topic in a data field or select from a previously entered component/ topic in the listbox above the data field. The purpose of this task is to consolidate all statements from a larger body of freeform text into primary themes focused on related system components/ topics of interest.

2.3.2 Classify Findings

The last step for the user is to assign an applicable generalizable HFE design principle that applies to the statement at hand. A three level hierarchical drill-down tree is provided to the user to select from. For this preliminary tool, 60 HFE design principles were selected from Wickens and colleagues [27]. One motivation for using this resource was that the book's primary purpose is to provide HFE practitioners with actionable HFE principles that can be used for system design. These design principles are all tied to NUREG-0700 guidelines, which are meant to be suggested guidelines that can mitigate potential HEDs for a given design issue. Statements with multiple system components/ topics or applicable HFE design principles can be run multiple times through the tool by selecting 'Repeat >' from the bottom menu. Additionally, the user can get detailed descriptions of each HFE design principle by selecting 'Get Info' from the bottom menu. If a statement is misclassified, the user is able to return to the previous statement by selecting 'Go Back' from the bottom menu.

2.4 Design Guidance

When the user classifies all statements from the HFE Issue Classification GUI, the tool will provide a list of all applicable NUREG-0700 guidelines for each HFE design principle with the listed system components/ topics. Fig. 4 illustrates an example of applicable NUREG-0700 guidelines to address the HMI button size design issue identified by the user. While not all guidelines provided from the tool may be appropriate, the tool greatly reduces the workload of systematically reviewing and selecting all 2195 guidelines.

x	x				
Human Factors Principle violated: Touch Zones	Component(s): HMI Button				
Level.1	Level.2	Level.3	Guideline. Number	Guideline.Title	Description
2 - USER-INTERFACE INTERACTION AND MANAGEMENT	User Input Formats	Direct Manipulation	2.2.6-10	Size of Icons	Items on the screen that are displayed for selection should be a minimum of 0.2 inch (5 millimetres) on a side and separated by at least 0.1 inch (3 millimetres).
2 - USER-INTERFACE INTERACTION AND MANAGEMENT	User Input Formats	Direct Manipulation	2.2.6-11	Text Selection Area	When functions are represented by text labels, a large area for pointing should be provided, including the area of the displayed label, plus a half-character distance around the label.
2 - USER-INTERFACE INTERACTION AND MANAGEMENT	User Input Formats	Direct Manipulation	2.2.6-12	Zooming for Precise Positioning	When data entry requires exact placement of graphic elements, users should be allowed to request expansion of the critical display area ("zooming") to make the positioning task easier.
2 - USER-INTERFACE INTERACTION AND MANAGEMENT	Cursors	Movement	2.3.3-6	Variable Step Size	When character size is variable, the incremental cursor positioning should vary correspondingly, with a step size matching the size of currently selected characters.
3 - CONTROLS	Input Devices	Touch Screens, Light Pens, and Graphic Tablets	3.2.4-10	Dimensions and Separation of Touch Zones	To allow for finger size and parallax inaccuracy, the dimensions of response areas of touch screens should be a maximum height and width of 1.5 inches (40 mm) and a minimum height and width of 0.6 inches (15 mm), with a maximum separation distance of 0.25 inches (6 mm) and minimum of 0.1 inches (3 mm).
3 - CONTROLS	Input Devices	Touch Screens, Light Pens, and Graphic Tablets	3.2.4-14	Light Pen Dimensions and Mounting	The light pen should be between 4.75 to 7 inches (120 to 180 mm) long with a diameter of 0.3 to 0.75 inches (7 to 20 mm). A conveniently located clip should be provided to hold the pen when not in use.
3 - CONTROLS	Input Devices	Touch Screens, Light Pens, and Graphic Tablets	3.2.4-15	Graphic Tablet Size and Orientation	Transparent grids that are used as display overlays should conform to the size of the display. Grids that are displaced from the display should approximate the display size and should be mounted below the display in an orientation to preserve directional relationships to the maximum extent.
2 - USER-INTERFACE INTERACTION AND MANAGEMENT	User Input Formats	Selection of Menu Options	2.2.2.5-10	Large Pointing Area for Option Selection	If menu selection is accomplished by pointing, the acceptable area for pointing should be as large as consistently possible, including at least the area of the displayed option label plus a half-character distance around that label.
3 - CONTROLS	General Control Guidelines	Coding of Controls	3.1.2-2	Size Coding Levels	No more than three different sizes of controls should be used for discrimination by absolute size.

Figure 4. Generated List of NUREG-0700 Guidelines.

3 PRELIMINARY EVALUATION OF THE TOOL'S PERFORMANCE

A preliminary evaluation of the tool's accuracy with classifying potential design issue was completed with a HFE qualitative dataset that was never exposed to the tool. The primary focus of this evaluation was to evaluate the number of missed statements that contained an actual design issue, which were classified as a non-issue by the tool. Additionally, comparing the inter-rater reliability of the tool to a baseline coding compared to a HFE expert's coding was evaluated. The dataset that was used was from a prior HFE study of an analog-to-digital migration of a Turbine Control System. The freeform notes from the original dataset were reformatted (i.e., with comments in a single comment) so that the tool could read the data. The content itself was not modified from the original dataset.

3.1 Method and Analyses

A qualitative dataset from a prior HFE study was used to evaluate the accuracy of the HFE Issue Classification tool compared to how HFE experts would classify HFE design issues. A total of 63 statements were provided in the dataset. Of these 63 statements, 38 were classified as HFE design issues where as 25 were classified as non-issues. The HFE expert who originally recorded the notes from the HFE study served as a baseline for comparing the tool to an independent HFE expert. The independent HFE expert was not involved in the original HFE study, and independently classified each statement as being a HFE design issue or non-issue. The HFE expert did not have any prior knowledge of how the original dataset was coded.

Performance measures of interest used to compare the tool to the independent HFE expert included frequencies of *hits* (true positives), *correct rejections* (true negatives), *false alarms* (false positives) and *misses* (false negatives). Other performance measures used included the *accuracy, kappa, precision, recall,* and the *F-measure*. Accuracy is a measure of success rate, accounting for the total correctly classified statements over total possible statements. The kappa statistic is a measure of accuracy with adjusting for the possibility of correctly classifying a statement by chance alone. Accuracy and kappa range from 0 to 1 where 1 refers to complete agreement and 0 refers to no agreement. Precision, or sensitivity, is the proportion of hits over the total number hits and false alarms. Recall, or specificity, is the proportion of hits over the total number of hits and misses. While precision captures the degree of classifying potential HFE design issues from noise, recall captures the degree of 'completeness' of identifying all HFE design issues. The precision and recall statistics range from 0 to 1 where a value of 1 refers to maximum precision or recall. Finally, the F-measure was used as an aggregate measure of performance to compare the tool to the independent HFE expert. The F-measure accounts for both precision and recall, using the harmonic mean, where values range from 0 through 1. Values closest to 1 denote more optimal overall classification performance.

3.2 Preliminary Results

Table I below presents the preliminary findings, comparing the tool's performance to an independent HFE expert.

Table I. Summary of classification performance: Comparison of the tool to an independent HFE expert.

Statistics	Tool	Independent HFE Expert
Hits	29	24
Correct Rejections	16	23
False Alarms	9	2
Misses	9	14
Accuracy	0.71	0.75
Kappa	0.40	0.51
Precision	0.76	0.92
Recall	0.76	0.63
F-measure	0.76	0.75

Overall, the tool performed similar to the independent HFE expert as shown from the F-measure. As expected, the tool was slightly more liberal with classifying potential HFE design issues, as reflected from a greater hit rate, greater false alarm rate, lesser correct rejection rate, lesser miss rate, and greater recall. However, this liberal classification approach by the tool did reduce the overall accuracy, kappa, and precision. Collectively from this small dataset, these preliminary results suggest that the tool may be more aggressive at classifying statements as potential HFE design issues, with the expense of having more false alarms. It should be emphasized that these results are more ideal than having a conservative approach with classifying potential HFE design issues, as the cost of missing a statement may mean missing important insights from the HFE T&E.

3.3 Limitations and Future Directions

While these results are promising for the development of this tool, there are limitations to this evaluation worth addressing. First, this preliminary evaluation used a relatively small single dataset. As such, the notes captured from these statements may not fully reflect all possible statements with potential HFE design issues concerning CRM upgrades in NPPs. Similarly, a single HFE expert was used to compare performance, which presents limitations to the validity of making inferences that the tool performs similar or better than HFE experts. Indeed, additional evaluations using more datasets as they become available with more HFE experts can strengthen the conclusions being made of this tool's performance classifying potential HFE design issues.

There are also areas to improve the tool. For one, there are other ML approaches from Naïve Bayes that may be used such as Random Forest, Support Vector Machines, or some combination of these algorithms using ensemble agreement. Secondly, building a larger and more diverse training dataset can improve classification performance. Third, consideration of automating the process of mapping identified potential HFE design issues to specific system component/ features and generalizable HFE design principles should be explored for an improved user experience. Fourth, the selection of generalizable HFE design principles and linking of them to NUREG-0700 guidelines was completed by a single HFE practitioner. Additional feedback to see if these design principles are clear, comprehensive, and can be reliably mapped to their appropriate NUERG-0700 guidelines should be pursued to improve the utility of this tool to real world application. Finally, there are other design guidance documents that may be applicable to CRM efforts to which the tool could include for a more comprehensive set of

recommendations. For example, there are various ANSI and ISO standards that are specific to HFE and HMI design, which may be more applicable to CRM applications.

4 CONCLUSIONS

This paper presents a novel tool that is intended to improve the quality of analysis completed by HFE practitioners in supporting CRM efforts funded by the DOE LWRS II&C System Technologies pathway. Specifically, the tool supports the analysis of qualitative freeform notes collected from T&Es during the HSI Design phase to readily provide actionable design guidance that can be traceable to state-of-the-art HFE guidelines such as NUREG-0700. Consequentially, such guidance provided by the tool should enable greater confidence in successfully transitioning from HSI Design into V&V. Further, this preliminary evaluation of the tools performance, compared to a HFE expert, showed similar classification performance for identifying potential HFE design issues. These results are promising and support continuation of exploring new ways to refine this tool so that it can be used in real world CRM applications.

5 ACKNOWLEDGMENTS

INL is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517. This work of authorship was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately-owned rights. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof. The INL issued document number for this paper is: INL/CON-16-40406.

6 REFERENCES

1. EIA, U. (2016). Energy Information Administration (2016), *Annual Energy Outlook 2014 with Projections to 2040, p. MT-15*. DOE/EIA-0383(2016).
2. U.S. Nuclear Regulatory Commission. (2007). *Standard Review Plan for the Review of Safety Analysis Reports for Nuclear Power Plants: LWR Edition—Human Factors Engineering, Chapter 18, NUREG-0800.* Washington, DC: U.S. Nuclear Regulatory Commission.
3. Bruseberg, A. (2008). Presenting the value of Human Factors Integration: guidance, arguments and evidence. *Cognition, Technology & Work, 10*(3), 181-189.
4. Boring, R. L., Ulrich, T. A., Joe, J. C., & Lew, R. T. (2015). Guideline for Operational Nuclear Usability and Knowledge Elicitation (GONUKE). *Procedia Manufacturing, 3*, 1327–1334.
5. Sauro, J. (2014). The relationship between problem frequency and problem severity in usability evaluations. *Journal of Usability Studies, 10*(1), 17-25.
6. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/.*
7. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.4.

8. Hennig, C. (2015). fpc: Flexible Procedures for Clustering. R package version 2.1-10. URL *http://CRAN.R-project.org/package=fpc.*

9. Ooms, J. (2016). hunspell: Morphological Analysis and Spell Checker for R. R package version 1.4.3. URL *http://CRAN.R-project.org/package=hunspell.*

10. Michalke, M. (2016). koRpus: An R Package for Text Analysis (Version 0.06-5). URL *http://reaktanz.de/?c=hacking&s=koRpus.*

11. Wild, F. (2015). lsa: Latent Semantic Analysis. R package version 0.73.1. URL *http://CRAN.R-project.org/package=lsa.*

12. Hornik, K. (2016). NLP: Natural Language Processing Infrastructure. R package version 0.1-9. URL *http://CRAN.R-project.org/package=NLP.*

13. Hornik, K. (2016). openNLP: Apache OpenNLP Tools Interface. R package version 0.2-6. URL *http://CRAN.R-project.org/package=openNLP.*

14. Rinker, T. W. (2013). qdap: Quantitative Discourse Analysis Package. 2.2.5. University at Buffalo. Buffalo, New York. URL *http://github.com/trinker/qdap.*

15. Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1-20.

16. Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1. URL *http://CRAN.R-project.org/package=SnowballC.*

17. Wickham, H. (2015). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0. URL *http://CRAN.R-project.org/package=stringr.*

18. Grosjean, P. (2015). SciViews: A GUI API for R. UMONS, Mons, Belgium. URL *http://www.sciviews.org/SciViews-R.*

19. Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2. URL *http://CRAN.R-project.org/package=tm.*

20. Mirai Solutions GmbH (2015). XLConnect: Excel Connector for R. R package version 0.2-11. URL *http://CRAN.R-project.org/package=XLConnect.*

21. OpenNLP, A. (2011). Apache software foundation. URL *http://opennlp.apache.org*.

22. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

23. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

24. Schneider, K. M. (2003, April). A comparison of event models for Naive Bayes anti-spam e-mail filtering. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (pp. 307-314). Association for Computational Linguistics.

25. Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning, 29*(2-3), 103-130.

26. Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.

27. Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (2003). *An introduction to human factors engineering (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.