2015 Asia-Pacific Conference on Computer Aided System Engineering

Mirroring Teachers' Assessment of Novice Students' Presentations Through an Intelligent Tutor System

Vanessa Echeverría
and Bruno Guamán
Information Technologies Center
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador
Email:vecheverria@cti.espol.edu.ec
Email:bguaman@cti.espol.edu.ec

Katherine Chiluiza
Faculty of Electrical and Computer Engineering
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador
Email:kchilui@espol.edu.ec

***Abstract*—This study proposes an Intelligent Tutor System for assessing slide presentations from novice undergraduate students. To develop such system, two learner models (rule based model and clustering model) were built using 80 presentations graded by three human experts. An experiment to determine the best learner model and students' perception was carried out using 51 presentations uploaded by students. The findings show that the clustering model classified in a similar way as a human evaluator only when a holistic evaluation criterion was used. Whereas, the rule-base model was more precise when the evaluation rules were easier to be followed by a human evaluator. Furthermore, students agreed with the usefulness of the system as well as the level of agreement with the grading model, although the latter in a lesser extent. Results from this study encourage to explore this area and adapt the proposed Intelligent Tutor System to other existing automated grading systems.**

***Keywords*-intelligent tutor system, automatic assessment, slide assessment, machine learning classifiers.**

I. Introduction

Oral presentations are part of the common tasks that students engaged in when they are at university. Slide presentations are regularly used as visual supporting materials in lectures and speeches [1]. These materials are used to obtain a more structured, attractive and interesting performance, which in turn would increase the amount of information that the audience comprehends [2]. Consequently, special care has to be taken when designing a presentation because its aim is to support and complement the speaker's discourse [3].

Common guidelines related to the design of slides for oral presentations have been identified in the literature. In [4] [5] [6], authors agreed on paying attention to the following aspects: font size and font family; contrast; images and distribution of the information, among others. They recommend that the font size should be larger than 18 points and smaller than 40 points, font families should be limited to two, preferring Sans Serif typeface. For the slide contrast, these authors suggest that presenters should avoid strident contrast, e.g. red-green, color combinations between foreground and background that would affect the readability of the slide. The images should be relevant and distributed accordingly, knowing that one slide with three or more images is less legible than two slides with one image each. Additionally, some advice about the distribution of the information to avoid too much text were given, such as: the information presented on each slide should not have more than six lines of text and seven words per line; each slide should have a headline; the information included should be important and representative to the topic; and, white space should be used generously.

Another guideline signaled by [3] indicated that presenters should be aware about the graphic design of presentations. For instance: avoiding templates or low-quality line art, using slide transitions, video and audio when it is strictly necessary and using high-quality graphics including photographs.

Although the above recommendations are available and referred by educators to novice undergraduate presenters, many students do not follow them, because either they do not know them, or they fail to interpret them correctly. Moreover, giving a personalized early feedback to students to support them in enhancing the quality of their presentations, is a time demanding task that educators struggle to cope with. Intelligent tutor systems (ITS) could be of help in this matter for teachers as well as novice students. An ITS is a system that learns a model from a specific domain and gives on-time feedback through the learned model without the advice of an expert [7].

The present research aims to answer the following question: how close an Intelligent Tutor System (ITS) mirrors teachers' grading of students' slide presentations? More specifically, this study compares the human assessment with two automatic grading models: one based on rules and the other model based on clustering. Both models are built-in the proposed ITS. Besides, the level of agreement about the grading offered by this system, as well as its usefulness

978-1-4799-7588-4/15 $31.00 © 2015 IEEE
DOI 10.1109/APCASE.2015.53

from students' perspectives are explored. The findings of this work indicate that the grading of the clustering based model nearly mirrored the grading of humans only when the criterion used is more holistic. In addition, students, who used the system, reported positive perspectives about their level of agreement with the proposed system's grading and its usefulness.

This study is structured as follows: section 2 includes related work about ITS for presentations; in section 3, the proposed solution is described along with its implementation; in section 4, the system validation is explained. The paper finishes with a discussion and further work section.

II. Related Work

A well-researched area, inside the learning technology field, is the one that studies ITS [8]. ITS can learn a model from any knowledge domain and help users in the learning process [7]. Several studies on ITS in a variety of knowledge areas have been reported in the research community [9] [10]. In these studies, it is reported that students reach a level of knowledge through suggestions or defined steps to learn algebra, mathematics and computer-related topics. On the same ITS area, the work of [11] reported a system that assessed and gave feedback to presenters. Authors implemented an intelligent tutoring system based upon an empirical research by extracting non-verbal behaviors from presenters and giving immediate feedback of their performance through a virtual room. In the same context, Batrinca et al. [12] exposed in their research, a platform to train presenters by analyzing the video, audio and depth camera recordings of the presenter, and giving online feedback interacting with agents inside a virtual reality environment.

Another study by Pattanasri et al. [13] was centered on estimating the slide comprehension using content-based features and presentation-based features. To extract content-based features, authors used natural language processing to obtain the main topics of the presentations. Likewise, to extract presentation-based features, they wrote down the information manually such as font size, bullets and charts. The final product of the study was an automated model obtained from students' questionnaire feedback.

The automatic evaluation of the quality of slide presentations, is a research area that has been recently explored. This is the case of [14], who matched the extracted features of slide presentation's with human evaluation, using a classifier. The classifier used features that can be automatically extracted such as: font size, number of words, images, charts and the image entropy of each slide for measuring the contrast. Furthermore, results reported an accuracy of 0.65 between the model and human evaluation. In addition, the study of Kim et al. [15] evaluated the quality of slide presentations based on information quality of slides represented into five dimensions. Furthermore, a model was created using 28 automatic extracted features such as: informativeness, cohesiveness, readability, ease of navigation and representational clarity. The results of this study showed that the precision of the model was 0.622 with respect to 200 presentations obtained from SlideShare and manually tagged by human annotators. Despite the interest in ITS for slide presentations, none of the reviewed studies give online feedback to novice students about the design choices they use in their slides.

The present study is complementary to systems like the ones proposed by [11] [12], that evaluated oral presentation skills of the presenter but did not take into consideration the slide presentations. An automated intelligent tutor that assesses the design of the slide presentations could be integrated into these systems to provide a holistic feedback about all the presentation aspects.

III. Proposed Solution

The proposed solution implements a web-based system that evaluates a slide presentation file, uploaded by a student. The system shows some recommendations in case the presentation file needs design improvements. Based on the ITS definition of [7], the system is composed by three models: the domain model, the learner model and the teaching model. In the following subsections these models are described.

A. Domain Model

The domain model aims to capture information of slide presentations concerning its design and related aspects. Here, a dataset composed of 80 presentations (.pptx file format), from a variety of topics and from undergraduate students, were analyzed to represent the specific domain. Furthermore, human evaluation was performed to generate the ground truth on which the proposed models should be compared.

1) Human Evaluation: Three human experts were asked to assess each presentation using a four-point grading category, being 4 the higher and 1 the lower. These categories were used to measure three criteria: amount of text, readability and contrast. The evaluation was performed under two conditions: a global condition for each slide presentation and an individual condition for three randomly chosen slides per presentation file. Next, inter-rater reliability coefficients were calculated from the human evaluations for both conditions. The Krippendorff's alpha for the global slide evaluations was 0.736, and the reliability coefficient for the individual slide evaluation was of 0.801. These coefficients are indicators of good reliability of human evaluators. The mode of the human evaluation per criterion was considered as the ground truth. Table I describes the frequencies of each grading category per criterion and condition.

2) Feature Extraction: Following the approaches described in [13] [14] [15], this research adopted most of the features mentioned in these studies.

Table I
PERCENTAGE OF FREQUENCIES OF PRESENTATIONS PER GRADING CATEGORY AND EVALUATION CRITERIA

	Global Presentations			Individual slides		
	Criteria			Criteria		
Grading Category	AT[1]	RD[2]	CNTR[3]	AT[1]	RD[2]	CNTR[3]
4	20.00	8.75	56.25	69.41	59.63	86.24
3	66.25	76.25	36.25	16.47	27.52	11.93
2	11.25	15	7.50	7.06	12.84	0.92
1	2.50	0	0	7.06	0	0.92

[1] AT=Amount of Text; [2] RD=Readability; [3] CNTR=Contrast

The automatic process of feature extraction used XSLF API built-in from Apache POI [1], which is a Java library for handling various file formats based on the Open Office XML Standards (OOXML). This API divides the content of a slide into any type of shape (e.g. TextShape, AutoShape, TableShape and PictureShape), where the text, colors, and other characteristics can be extracted. Consequently, features related to each evaluation criteria, which are described below, were determined by using the aforementioned process.

- *Amount of text (AT):* The number of lines *(NL)* and words *(NW)* is retrieved from each TextShape.
- *Readability (RD):* Each TextShape is divided into TextRuns, which are text fragments differentiated from others by their font formats. Font family *(FF)* and font size *(FS)* characteristics are extracted.
- *Contrast (CONT):* As defined in WCAG 2.0 [16], the intent of this characteristic is to provide a good color combination of text and background. The measurement for identifying a good contrast is defined by the Contrast Ratio (*CR*), which is calculated by the following formula: $CR = (L1 + 0.05)/(L2 + 0.05)$, where *L1* is the relative luminance (*RL*) of the lighter of the colors, and *L2* is the *RL* of the darkest color.Thus, to calculate the contrast of a slide, the colors for text and for background are extracted for each TextShape. The predominant color is obtained from the background color of the TextShape by using ColorWave color clustering [2] due to the existence of diverse colors at presentations' background. At the end, the luminance of the text color and the luminance of the background color are used to calculate the *CR* for each TextShape. The range of values obtained from using the CR formula are between 1 to 21. A value of 21 reflects a very good contrast. Three Contrast levels (*CL*) were defined, depending on the *CR* values. Contrast level A is the lowest (CR <4.5); contrast level AA is greater of equal to 4.5 and less than 7.0 and contrast level AAA is the highest (CR >=7.0).

B. Learner Model

The learner model aims to build the automatic evaluation through the adoption of any machine learning technique. Machine learning techniques bring to intelligent tutor systems, different models and classifiers to perform automatic learning. However, an appropriate model selection that fits the problem and data is needed. Thus, the selection of the learner model was carried out by comparing two models.

The first model was constructed using specific rules taken from the literature [4] [5] [6], whereas the second model was created using K-means clustering. K-means clustering was preferred from other machine learning models due to the characteristic of generating representative groups of data according to the selected number of clusters.

1) Rule based model: A set of rules was created according to the three evaluation criteria and conditions (global and individual), which are explained below:

Amount of text: For the individual condition, the number of lines in a slide *(NLS)* (see section III-A2) should be six or less, whereas the proportion of slides with six or less number of lines was the rule for the global condition.

Readability: For the individual condition, the font size should be greater than 18 points in a slide (section III-A2), while the proportion of slides with a font size greater than 18 points was the rule for the global condition.

Contrast: For the individual condition, the contrast level in a slide should be AAA (see section III-A2), while the proportion of slides that fulfilled the AAA level contrast was the rule for the global rule.

Therefore, the rule based model ended up with six rules. Once the rules were implemented, they were tested using the human evaluation criteria. Table III shows the accuracy results of the testing per criterion.

2) Clustering-based Model: Six clustering models were built based on each criterion and condition (global and individual). A feature selection was performed to find relevant features for each criteria, separately. As can been seen in Table II, relevant features for the amount of text, readability and contrast are: number of words and number of lines; minimum font size; and contrast level for each slide. Moreover, the maximum, average and standard deviation of the number of words; the minimum argument of the minimum font size and the contrast level, were selected as relevant features for the global condition.

Due to the fact that there were unbalanced grading categories (See table I), categories *1* and *2* were merged, ending up with three grading categories. Thus, data were clustered and evaluated using human scores. Table III shows the accuracy for each clustering model.

C. Teaching Model

The purpose of the teaching model is to evaluate automatically a presentation and to give feedback to users through recommendations and the grading category obtained from

[1] https://poi.apache.org
[2] https://pypi.python.org/pypi/colorweave/0.1

Table II
INDIVIDUAL AND GLOBAL FEATURES USED TO BUILD CLUSTERING MODELS PER CRITERION AND CONDITION

Criterion	Condition	Feature	Description
Amount of Text	Individual	numberWords (*NWS*) numLines (*NLS*)	Number of words in a slide Number of lines in a slide
	Global	maxNumberWords (*MAXNWP*) avgNumberWords (*AVGNWP*) stdNumberWords (*STDNWP*)	Maximum of the *NWS* values in the presentation Average of the *NWS* values in the presentation Standard deviation of the *NWS* values in the presentation
Readability	Individual	minFontSize (*MINFSS*)	Minimum font size in a slide
	Global	minMinFontSize (*MINMINFSP*)	Minimum of *MINFSS* values along the presentation
Contrast Ratio	Individual	contrastLevel (*CLS*)	WCAG 2.0 contrast level, based on the mode of *CR* values in a slide *MODCRS*
	Global	contrastLevelP (*CLP*)	WCAG 2.0 contrast level, based on the mode of *MODCRS* values along the presentation.

Table III
ACCURACY FOR GRADING MODELS

Condition	Criterion	Clustering Accuracy	Rule based Accuracy
Global	Amount of Text	0.6049	0.3827
	Readability	0.5625	0.4814
	Contrast	0.6145	0.2469
Individual	Amount of Text	0.4942	0.6104
	Readability	0.4495	0.6104
	Contrast	0.6273	0.4815

the learner model. To accomplish the purpose of this model, a web-based system was developed in Django Framework for Python[3]. Figure 1 depicts the architecture of the system. Django is the baseline of the web-application, while the extraction of the features and clustering task are performed in Java with the Apache POI API and Weka API [17], respectively. Java tasks communicate with Python tasks using Pyjnius wrapper[4]. The web-based system works as follows: first, from an uploaded presentation, the system extracts the features according to the approach presented in section III-A2. Then, the learner model evaluates the presentation using extracted features and obtains individual and global grading categories per criterion. Finally, recommendations are established based upon these grading categories regardless of the learner model.

IV. SYSTEM VALIDATION

Third-year undergraduate students from an engineering oriented university prepared fifty-one presentations that were used to validate the two learner models. These presentations were assessed later by an expert in evaluation of presentations. In addition, students that provided the presentation files were asked to respond to a short survey about their level of agreement with the grading and feedback offered by the ITS, as well as its perceived usefulness. The level of agreement was measured using a Likert scale, being 1, totally disagree and 5 totally agree; as for the usefulness of the proposed ITS, it was measured using a Likert scale (1 totally useless and 5 totally useful).

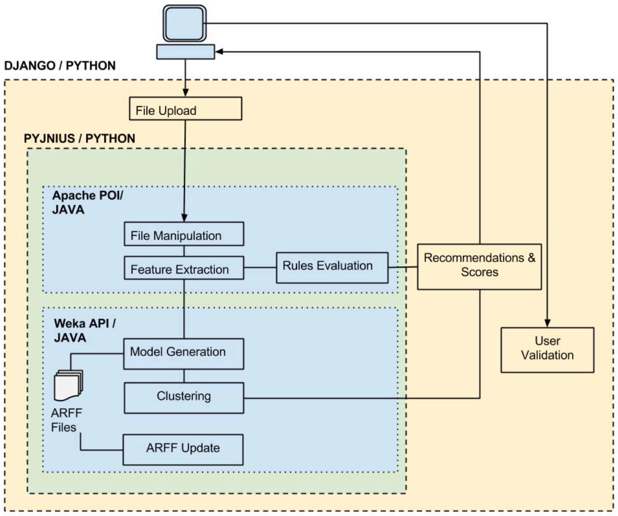

Figure 1. Architecture of the implemented intelligent tutor system.

A. Learner Models Testing

The outcomes from each learner model were contrasted with the human assessment. Calculations of precision and recall for each evaluation criteria per category for both learners models are included in Table IV. In general, very low average values for precision and recall are observed in both models with the exception of precision for Contrast in the rule based model (0.59) and precision for Amount of text in the clustering model (0.62). Looking at the level of categories, there are some values above 0.5. In the rule based model, the values for recall in category 3 for contrast and readability are greater than 0.6; similarly, precision in

[3] https://www.djangoproject.com
[4] https://pyjnius.readthedocs.org

Table IV
Precision and Recall Indexes for each Evaluation Criteria per Learner Model

				RULE BASED MODEL						**CLUSTERING MODEL**					
	Contrast	**Readability**	**Amount of Text**	**Contrast**		**Readability**		**Amount of Text**		**Contrast**		**Readability**		**Amount of Text**	
Cat.[1]	**Percentage Frequency**			**Pr**	**Recall**	**Pr**	**Recall**	**Pr**	**Recall**	**Pr**	**Recall**	**Pr**	**Recall**	**Pr**	**Recall**
1	3.92	5.88	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	13.73	15.69	0	0	0	0	0	0	0	0	0	0	**0.25**	**0.31**
3	23.53	39.22	43.14	**0.48**	**0.75**	**0.69**	**0.67**	0.60	0.18	0.27	0.30	0.57	0.80	**0.82**	**0.53**
4	37.25	3.92	5.88	**0.70**	**0.24**	**0.19**	**0.80**	0.07	0.50	0.54	0.24	0.17	0.20	**0.15**	**0.50**
	Average			**0.59**	0.43	**0.42**	0.47	0.41	0.16	0.41	0.25	0.35	0.49	**0.62**	0.47

[1] Cat=Category

category 4 is even higher (0.7). The same applies for recall in the category 4 for Readability (0.8). As for the clustering model, the precision value for category 3 is the highest observed in the table. The percentage of frequencies per category are also included in this table. It is obvious that there are negative skewed distributions for each evaluation criteria; there are few observations in categories 1 and 2, whereas categories 3 and 4 include more observations.

B. *Students' perspectives about the proposed ITS*

Overall students reported positive perspectives related to the ITS (median for both variables were 4). Both distributions were negatively skewed (level of agreement= -0.487, usefulness=-0.685). Note from Figure 2 that the perceived usefulness of the proposed ITS was higher than the levels of agreement related to grading of ITS.

V. Discussion and Future Work

The central research question of this study was: How close an ITS mirrors teachers' grading of students' slide presentations? To answer this question two learning models based on rules and clustering were tested. The results show in general that neither the rule based model, nor the clustering model are good enough in comparison to a human evaluator. Despite the grading models used to build the classifiers showed better accuracy, the results when testing the models were challenging. Only the clustering model achieved an average precision of 0.62 for the amount of text criterion. Likewise, the rule based model showed a precision of 0.59 for the contrast criterion, neither of the models show an important indicator of precision for readability.

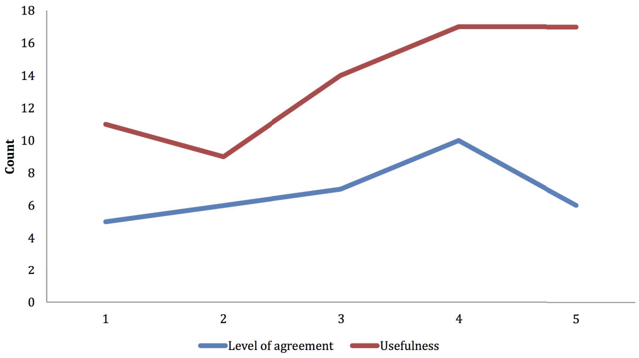

Figure 2. Students' Perceptions about Level of Agreement with Grading and Usefulness of the Proposed ITS

As for the learner models, on one hand, the rule based model seemed to have a better performance for contrast and readability criteria. Again, the precision and recall values are better for categories 3 and 4. The rules applied for contrast and readability seem to go in line with the perceptions of human evaluators. Perhaps, the rules associated to these criteria were easier to be followed, whereas the set of rules for amount of text rules seem not that natural to be applied for a human being. In other words, it is unnatural for a person to count the number of lines that a slide presentation includes; a person perceives an adequate or inadequate number of lines in a slide and according to this, he/she grades the presentation. On the other hand, the clustering model achieved better indicators, precisely in the amount of text criterion. Clustering is built upon the features where more homogeneity is observed, which are the final clusters. Human beings do not only grade based upon one specific characteristic of a given criteria. Human evaluators use several characteristics in a holistic way [18] [19].

It is important to note that both, the grading model and learner models, used skewed distributions in all the categories (more observations in categories 3 and 4). The lack of enough observations for categories 1 and 2 jeopardized the precision and accuracy of the classification models. The clustering model was built with a fixed number of clusters (categories 1 to 4); therefore, if no or few observations were included in the lower categories (1 and 2), the model could erratically classified them.

On the side of the students, they found that the proposed system is a very useful tool, which is in line to the findings in [20], about the perceived usefulness of automatic assessment tools. Novice students, as well as any lecturer sometimes struggle when dealing with design aspects [21], which in the particular case of slide presentations, are key for successful presentations. Thus, a system that provides specific suggestions at the granularity of individual slides is desirable. Despite students reported to a lesser extent their agreement with the grades offered by the proposed ITS, overall their perceptions on this variable were positive. This finding is not unexpected, human beings are skeptical about receiving grades or feedback from a computer [22] [23].

A next natural step for enhancing the proposed system could be to input the presentations that coincided in grading score with the human evaluator into the system. This action could guarantee active on-line learning and perhaps a better learner model of the proposed ITS. Even though the outcomes of this study are still challenging, more work about this work is foreseeing, as well as more research related to the inclusion of non-text elements for presentations' assessment. Learner models based on fuzzy logic should be explored due to the holistic approach followed by humans when evaluating presentations. The semantic analysis of the content of presentations is another aspect that needs to be considered in a future research agenda.

ACKNOWLEDGMENT

The authors would like to thank SENESCYT for their funding support to the Andamios project.

REFERENCES

[1] N. Erdemir, "The effect of powerpoint and traditional lectures on students' achievement in physics," *Journal of Turkish Science Education*, vol. 8, no. 3, pp. 176–189, 2011.

[2] M. Carter, *Designing science presentations: a visual guide to figures, papers, slides, posters, and more*, 2012.

[3] G. Reynolds, *Presentation zen design: simple design principles and techniques to enhance your presentations*. New Riders, 2013.

[4] E. Cheney, "No more lousy powerpoint slides," *GSA Today*, vol. 23, no. 9, pp. 68–69, 2013.

[5] D. Brodsky and E. Doherty, "Creating an effective powerpoint presentation," *NeoReviews*, vol. 12, no. 12, pp. e687–e697, 2011.

[6] M. Alley and K. Neeley, "Rethinking the design of presentation slides: A case for sentence headlines and visual evidence," *Technical Communication*, vol. 52, no. 4, pp. 417–426, 2005.

[7] F. Akhras and J. Self, "Beyond intelligent tutoring systems: Situations, interactions, processes and affordances," *Instructional Science*, vol. 30, no. 1, pp. 1–30, 2002.

[8] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, "Intelligent tutoring goes to school in the big city," 1997.

[9] V. Aleven, B. M. Mclaren, J. Sewall, and K. R. Koedinger, "A new paradigm for intelligent tutoring systems: Example-tracing tutors," *International Journal of Artificial Intelligence in Education*, vol. 19, no. 2, pp. 105–154, 2009.

[10] A. Graesser, P. Chipman, B. Haynes, and A. Olney, "Auto tutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612–618, 2005.

[11] A.-T. Nguyen, W. Chen, and M. Rauterberg, "Mastering the art of persuasion: Intelligent tutoring system for presenters," vol. 3, 2014, pp. 83–90.

[12] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero - towards a multimodal virtual audience platform for public speaking training," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8108 LNAI, pp. 116–128, 2013.

[13] N. Pattanasri, M. Mukunoki, and M. Minoh, "Learning to estimate slide comprehension in classrooms with support vector machines," *Learning Technologies, IEEE Transactions on*, vol. 5, no. 1, pp. 52–61, First 2012.

[14] G. Luzardo, B. Guaman, K. Chiluiza, J. Castells, and X. Ochoa, "Estimation of presentations skills based on slides and audio features," 2014, pp. 37–44.

[15] S. Kim, W. Jung, K. Han, J.-G. Lee, and M. Yi, "Quality-based automatic classification for presentation slides," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8416 LNCS, pp. 638–643, 2014.

[16] B. Caldwell, M. Cooper, L. Guarino, G. Vanderheiden, P. Reutemann, and I. H. Witten, *Web Content Accessibility Guidelines (WCAG) 2.0*. W3C Recommendation, 2008.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, 2009, vol. 11.

[18] M. Zhang, "Contrasting automated and human scoring of essays," *R & D Connections*, vol. 21, 2013.

[19] M. D. Shermis and J. Burstein, *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

[20] H. D. Torsten REINERS, Carl DREHER, "Six key topics for automated assessment utilisation and acceptance," *Informatics in Education*, vol. 10, no. 1, pp. 47–64, 2011.

[21] R. J. Craig and J. H. Amernic, "Powerpoint presentation technology and the dynamics of teaching," *Innovative Higher Education*, vol. 31, no. 3, pp. 147–160, 2006.

[22] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[23] S. Tseng and B. Fogg, "Credibility and computing technology," *Communications of the ACM*, vol. 42, no. 5, pp. 39–44, 1999.